Title:

Towards Transparent and Interpretable Predictions of Student Performance Using Explainable AI

By

Samuel Forson Kwakye

A Thesis submitted in partial fulfilment of the requirement of the University of East London for the degree of Professional Doctorate in Data Science.

September 2024

Abstract

Artificial Intelligence (AI) is increasingly being adopted in educational contexts to support data-driven decision-making, particularly in predicting student outcomes. However, the opaque nature of many high-performing models raises concerns around fairness, accountability, and interpretability which are factors that are especially critical in high-stakes environments such as GCSE examinations. This study investigates how Explainable AI (XAI) techniques can enhance the transparency and interpretability of machine learning models used to predict GCSE English Language and Mathematics performance.

Using a real-world dataset from a secondary school in England, this research developed and evaluated predictive models, including Histogram-based Gradient Boosting (HGB) and a Multi-Layer Perceptron (MLP), to estimate student achievement outcomes. To address imbalances and maximise performance, the pipeline incorporated data pre-processing, feature engineering, and fairness-aware resampling strategies. The final HGB model achieved strong predictive accuracy while maintaining robustness across subgroups.

To ensure interpretability, four XAI techniques namely SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), PDP (Partial Dependence Plots), and ALE (Accumulated Local Effects) were applied. These methods provided insight into the most influential features driving predictions, including attendance, CAT3 scores, SEN status, and EAL. Novel explainability metrics such as transparency score, explainability ratio, and interpretability ratio were proposed to systematically evaluate explanation quality and model clarity.

In addition to technical evaluation, the study employed a stakeholder-centred design to assess how teachers, school leaders, and students interact with and interpret model explanations. Mixed-methods user studies revealed that personalised, context-sensitive explanations improved stakeholders' decision confidence, supported intervention planning, and prompted critical reflection. Concerns were also raised about fairness, overreliance, and the ethical implications of demographic profiling.

The research demonstrates that explainable models can enhance trust, transparency, and pedagogical utility when appropriately designed and evaluated in real-world educational settings. By integrating technical rigour with ethical and user-centred evaluation, this work

contributes to the development of responsible, interpretable AI systems that align with the values and needs of educators and learners. The study offers both methodological innovations and practical recommendations for the responsible deployment of XAI in education.

Ethical Approval

This project required ethical approval from the University of East London Ethics Committee, and confirmation of approval is under the application ID: ETH2223-0009 updated in September 2022. The ethics approval is provided in the Appendix section of this research work.

Acknowledgement

As Margaret J. Wheatley puts it, “Our willingness to acknowledge that we only see half the picture creates the conditions that make us more attractive to others. The more sincerely we acknowledge our need for their different insights and perspectives, the more they will be magnetized to join us”.

The subtle or obvious assistance from you all is the critical component that helped me complete this research work. This thesis is the culmination of years of research, learning, and support from many people, without whom this journey would not have been possible.

Therefore, I thank everyone who has crossed my path while undertaking this research work. I would like to express my sincere gratitude to Dr Nadeem Qazi, who has been with me through thick and thin, supporting me, guiding me and helping me to develop key research skills which have been invaluable throughout this research process. Your insightful feedback, valuable suggestions, and constant support have been instrumental in shaping this research, and I am deeply appreciative of the opportunities you have provided me.

I am grateful to my colleagues and friends at Davenant Foundation School who have been a source of motivation and inspiration. Special thanks to the Head teacher, Mr. Adam Thorne and the data manager for giving me access to data for this study.

A special thank you goes to my family Shirley, Jason and Alyssa for their unconditional love, understanding, and support throughout this long journey not forgetting Frank, Samil and Felix for your unwavering support and assistance.

Finally, to my friends outside academia, thank you for keeping me grounded and reminding me of the importance of balance. This thesis is dedicated to everyone who has contributed to my personal and professional growth during this journey. Thank you.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Full Form
AI	Artificial Intelligence
ANN	Artificial Neural Networks
ALE	Accumulated Local Effects
CAT	Cognitive Abilities Test
DL	Deep Learning
DNN	Deep Neural Networks
DT	Decision Tree
EGB	Extreme Gradient Boosting
ELM	Extreme Learning Machine
FCBF	Fast Correlation-Based Feature Selection
FSM	Free School Meals
GB	Gradient Boosting
HGB	Histogram-Based Gradient Boosting
KS2	Key Stage 2 (assessment level in UK primary education)
KS4	Key Stage 4 (assessment level in UK secondary education)
KNN	K-Nearest Neighbor
LIME	Local Interpretable Model-Agnostic Explanations
LSTM	Long Short-Term Memory
ML	Machine Learning
NB	Naïve Bayes
NN	Neural Network
PDP	Partial Dependence Plot
RF	Random Forest
SEN	Special Educational Needs
SHAP	Shapley Additive Explanations

SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
XAI	Explainable Artificial Intelligence

Chapter 1 – Research Background and Introduction

The integration of Artificial Intelligence (AI) in education is revolutionizing pedagogical principles, instructional methods, and learning experiences. As AI continues to reshape the educational landscape, the need for Explainable AI (XAI), which refers to a branch of AI that makes model predictions interpretable to humans, becomes increasingly evident. XAI addresses the critical demand for transparency in AI-driven decision-making, enabling educators, students, and stakeholders to understand how AI models generate predictions. This study explores contemporary research on XAI in education, examining its implications, applications, and challenges while positioning it within the broader context of AI adoption in learning environments. Specifically, this research focuses on adapting existing XAI models for predicting student performance and enhancing their interpretability to support informed decision-making in educational settings. This chapter introduces the foundational concepts of XAI in education, outlines the specific challenges faced in predictive grading, and presents the rationale, aims, and objectives of the study.

AI has demonstrated remarkable potential in personalized and adaptive learning, tailoring educational experiences to diverse learning styles and needs. By analysing large datasets, AI customizes learning pathways based on students' strengths and weaknesses in real time. Göçen & Aydemir (2020) emphasize AI's role in personalized education, particularly in adaptive learning systems, where real-time analytics enhance student engagement and learning outcomes. However, as AI becomes more embedded in education, concerns about trust, accountability, and ethics emerge. Popenici and Kerr (2017) discuss AI's impact on pedagogical roles, leading to intelligent tutoring systems and AI-assisted teaching models. While these innovations offer significant benefits, they also raise concerns about the transparency of AI decision-making. Hidayat et al. (2022) illustrate how AI enhances student achievement in mathematics but emphasize the need for explainable models to maintain trust and ensure effective educational outcomes.

Various XAI methodologies can improve transparency in AI-driven education. For example, Bayesian Teaching, though traditionally used in medical applications, provides insights into improving AI interpretability and fostering better human-AI interactions (Folke et al., 2021).

By understanding how AI models arrive at conclusions, educators can more effectively leverage AI tools to enhance student engagement and academic performance. AI-assisted teaching models also show promise in language education. Luo (2024) highlights AI-driven improvements in teaching English listening and speaking, shifting from static content delivery to interactive, personalized learning experiences. This transformation aligns with the broader shift in traditional education, where adaptive methodologies foster deeper student engagement. Beyond instruction, AI optimizes administrative tasks, improving educational efficiency through data-driven decision-making. Zheng and Badarch (2022) highlight AI's role in shaping institutional policies and automating administrative functions, allowing educators to focus more on student relationships and curriculum development.

Despite AI's benefits, its widespread adoption necessitates addressing ethical concerns. Chen et al. (2020) warn of biases in AI algorithms that could perpetuate inequalities in learning experiences. Transparency is crucial for ensuring fairness, inclusivity, and respect for all learners. Zhou (2024) argues that AI deployment must be accompanied by equitable resource distribution to ensure accessibility for students from diverse socio-economic backgrounds. Policymakers, educators, and technology developers must collaborate to construct robust, ethical AI-integrated educational ecosystems. Interdisciplinary collaboration is also essential. Wang and Xie (2024) stress the importance of academic engagement in shaping AI-driven educational technologies. By incorporating diverse perspectives, institutions can develop a comprehensive understanding of AI's role in education and optimize its implementation for meaningful learning experiences. As the discourse on AI in education evolves, it is crucial to continuously assess its applications and implications for educational reform. Zheng and Badarch (2022) emphasize that AI's effectiveness depends on its alignment with instructional methodologies and overarching educational goals. A critical evaluation of AI's role enables informed decision-making, ensuring that AI integration enhances both educational effectiveness and equity.

One of AI's most promising applications in education is student performance prediction. AI-driven models analyse vast datasets encompassing student behaviour, engagement, and academic history to forecast future performance. Niu et al. (2021) demonstrate that accurate prediction models enable proactive intervention, reducing the risk of academic failure. However, Amann et al. (2020) argue that for these predictions to be truly beneficial, stakeholders must understand how AI models generate insights. Explainability fosters trust in

AI systems, empowering educators to use AI-driven recommendations effectively while maintaining pedagogical authority. Recent studies, such as Ouyang et al. (2023), explore AI's integration with learning analytics in online education, highlighting how systematic AI modelling enhances learning outcomes. Similarly, Jobin et al. (2019) emphasize the importance of legal and ethical frameworks in AI adoption to ensure compliance with data privacy regulations and student rights. XAI enhances transparency in student performance prediction models, enabling more personalized learning interventions. Adnan et al. (2021) propose integrating machine learning with XAI techniques to help educators interpret students' academic trajectories. Improved prediction accuracy and interpretability empower educators to make informed decisions that enhance student learning experiences.

Ethical considerations remain at the forefront of AI-driven education. Shahzad et al. (2024) argue that AI must be employed equitably to support students from diverse backgrounds. Establishing AI-specific guidelines within existing educational frameworks ensures that AI serves as an equitable tool for improving access to education. Various AI-driven predictive models, including similarity-based, model-based, and probabilistic approaches, are being developed to analyse student engagement and learning patterns (Jiao et al., 2022). However, student privacy and data ethics remain critical concerns. Holmes et al. (2021) advocate for a community-wide framework to address the ethical dimensions of AI in education, ensuring transparency, accountability, and responsible data usage. The intersection of XAI and student performance prediction represents a significant evolution in educational methodologies, driven by data-informed decision-making. As educational institutions harness AI's potential, they must balance technological advancements with ethical considerations to promote transparency and fairness. Continued research, collaboration, and dialogue are essential to refining AI systems that align with pedagogical goals while maintaining institutional integrity.

As AI technologies advance, stakeholders must ensure their application is guided by transparency, fairness, and inclusivity. This thesis aims to explore the impact of XAI in education, specifically in monitoring and predicting student performance. It seeks to answer the question: How can XAI techniques enhance the transparency and interpretability of machine learning models used for student performance prediction?

By understanding the reasoning behind a model's predictions, educators and practitioners can build greater trust in AI-driven insights. This transparency not only fosters confidence in AI-powered educational tools but also encourages wider adoption, ultimately supporting more informed decision-making and improving student outcomes.

1.1 Identified Challenges in Adopting AI for Prediction in Education and the Need for Explainability

AI models in education offer substantial predictive capabilities but often lack explainability, a factor critical to trust and adoption. Two core concepts underpin explainability: transparency, which refers to how clearly the inner workings of an AI model can be understood (e.g., through model structure or documentation), and interpretability, which concerns how well a human especially a non-technical stakeholder can comprehend and act on the AI's outputs. Many AI models function as "black boxes," obscuring both transparency and interpretability. This section examines the multifaceted challenges spanning technical, ethical, psychological, and regulatory dimensions, reinforcing the need for XAI in educational contexts to mitigate these concerns.

1.1.1 Bias and Inequity in AI Predictions

One of the most pressing concerns is the potential for AI systems to perpetuate existing biases, resulting in inequitable outcomes. AI models trained on historical data may reflect societal and institutional inequalities, leading to unfair predictions for marginalized groups. For instance, models built on data from high-performing schools may undervalue students from under-resourced schools, reinforcing systemic disadvantages (Holmes et al., 2021).

This lack of fairness is further compounded by opacity. When stakeholders cannot interrogate how predictions are generated, addressing discriminatory outcomes becomes nearly impossible. XAI provides tools to detect and mitigate such biases by offering transparent explanations of AI decision-making.

1.1.2 Complexity of Educational Data

Educational settings are inherently complex. Students vary widely in terms of learning styles, contexts, and engagement levels. AI systems often fail to capture this diversity, especially when contextual data is lacking or ignored. This can result in inaccurate or misleading predictions that undermine student support strategies (Calatayud et al., 2021).

Explainability helps educators understand when and why AI models make errors, enabling them to adjust interventions accordingly. Without such insights, even high-accuracy models can be ineffective or harmful in practice.

1.1.3 Risk of Over-Reliance and Dehumanization

The increasing reliance on AI tools such as automated grading and behavior monitoring systems risks dehumanizing education. These systems may prioritize patterns over pedagogical judgment, failing to assess creativity, critical thinking, or context (Sullivan et al., 2023).

This over-reliance is dangerous when educators are unable to question or override AI outputs. XAI reasserts human agency by making the inner workings of AI transparent, allowing teachers to use AI as a supportive tool rather than a replacement.

1.1.4 Psychological Impact on Students

AI-generated predictions can influence students' self-perceptions and aspirations. When these predictions are opaque or overly deterministic, they may cause unnecessary anxiety or discourage students from pursuing certain goals. Research has shown that students often internalize negative AI feedback, even when inaccurate, which can lead to reduced motivation and disengagement (Selwyn, 2022).

XAI addresses this by making predictions more understandable and less intimidating. When students and educators can interpret the reasoning behind a prediction, they are better equipped to act on it constructively.

1.1.5 Lack of Stakeholder Understanding and Trust

Teachers, students, and school leaders often struggle to understand the rationale behind AI outputs, especially when models are complex and poorly documented (Manheim, 2019). This knowledge gap undermines trust and limits the effectiveness of AI tools in real-world classrooms.

XAI acts as a bridge between technical systems and human users. By offering clear, user-friendly explanations, it supports collaboration between AI and educational stakeholders, enhancing both understanding and adoption.

1.1.6 Regulatory and Philosophical Perspectives on Ethical AI

The adoption of AI in education is also shaped by an evolving regulatory landscape. The UK government's AI White Paper (2023) emphasizes a pro-innovation approach to AI regulation**,** underscoring the importance of transparency, accountability, and fairness in high-impact sectors like education. While not legally binding, the framework encourages sector-specific regulators such as Ofsted (Office for Standards in Education, Children's Services and Skills) and Ofqual (Office of Qualifications and Examinations Regulation) to establish clear guidance on the responsible deployment of AI technologies in schools.

Ofqual, in particular, has acknowledged the importance of transparency and fairness in algorithmic grading, especially in the aftermath of the 2020 grading controversy. Meanwhile, Ofsted has expressed interest in the potential of AI to support educational inspection and improvement, while cautioning against opaque decision-making systems that cannot be justified or understood by educators and families (Centre for Data Ethics and Innovation, 2020).

These regulatory efforts align with global ethical AI initiatives, including the EU Artificial Intelligence Act and UNESCO's Beijing Consensus, both of which highlight the need for explainable and equitable AI systems. The EU AI guidelines emphasize transparency and human oversight for high-risk AI applications such as education (European Commission, 2021), while UNESCO (2019) stresses human-centred approaches in AI deployment, ensuring technology supports and complements rather than replaces pedagogical judgement.

1.1.7 Ensuring Ethical, Inclusive, and Transparent AI Adoption

Beyond accuracy, AI tools in education must align with ethical values, especially in protecting student privacy and promoting equity. Without proper oversight, AI-based interventions may unintentionally reinforce stereotypes or exclude vulnerable learners (Khosravi et al., 2022).

XAI supports ethical AI development by exposing hidden patterns, allowing institutions to identify unintended harms, and fostering greater institutional accountability.

1.1.8 Making XAI Accessible to All Stakeholders

For XAI to be effective, its outputs must be understandable to diverse stakeholders. Teachers need pedagogically relevant explanations; students need motivational insights; policymakers require systemic overviews. One-size-fits-all explanations are insufficient.

This thesis addresses this need by advocating stakeholder-specific XAI strategies tailoring interpretability methods to the goals and expertise of each user group.

1.1.9 Summary

In summary, the widespread adoption of AI in education introduces serious challenges related to bias, opacity, trust, and psychological impact. XAI offers a viable solution to these problems by promoting transparency, enabling human oversight, and safeguarding fairness. Embedding XAI into educational AI systems is essential for developing tools that are accurate, ethical, and supportive of human-centred learning. As UK regulators and international frameworks increasingly emphasize explainability, this research aligns with broader societal efforts to ensure AI enhances, rather than compromises, equity and educational integrity.

1.2 Problem Statement and Research Gap

Many AI models function as "black boxes," offering little transparency or interpretability. This lack of explainability raises concerns regarding trustworthiness and effectiveness, as stakeholders such as educators, students, and administrators are often are unable to comprehend or act on AI-generated recommendations. Moreover, AI-driven predictions are susceptible to biases, particularly those influenced by socio-economic and demographic disparities, potentially reinforcing educational inequalities. Without mechanisms to elucidate AI decision-making, the educational sector risks misinterpretation of predictions, misclassification of students, and missed opportunities for timely intervention.

While XAI has been extensively explored in various other domains, its application in education particularly in student performance prediction remains underdeveloped. Few studies have examined the potential of XAI techniques such as SHAP (Shapley Additive explanations), LIME (Local Interpretable Model-agnostic Explanations), ALE (Accumulated Local Effects), and PDP (Partial Dependence Plots) in educational contexts, particularly in high-stakes examinations like the General Certificate of Secondary Education (GCSE). Existing literature underscores the necessity of transparency in AI-driven predictions to foster trust and usability in education (Ouyang et al., 2023). However, most AI models emphasize predictive accuracy while overlooking interpretability, failing to accommodate the diverse needs of learners and educators.

This gap is further compounded by the limited comparative research on multiple XAI techniques within a unified educational framework. The effectiveness of different XAI methodologies in student performance prediction remains largely unexplored (Livieris et al., 2018). Additionally, existing AI models often fail to incorporate critical student-specific factors such as learning styles and engagement levels (Trisnawati et al., 2023), leading to oversimplified assessments that do not account for the complexities of real-world learning environments.

Another critical but underexamined issue is the psychological impact of opaque AI predictions on students. Research indicates that students may experience heightened anxiety when confronted with AI-generated evaluations lacking clear explanations (Kim et al., 2022). Without transparency, AI-based assessments or predictions can diminish student confidence and engagement rather than providing constructive, actionable feedback.

Addressing these limitations requires not only technical innovation but also careful adaptation of XAI frameworks to educational realities. This research addresses these limitations by systematically evaluating SHAP, LIME, ALE, and PDP within a single framework for GCSE performance prediction. By integrating fairness constraints offered by the adapted techniques to mitigate socio-economic biases and introducing tailored explanations for different stakeholders, this study aims to enhance the transparency, interpretability, and ethical deployment of AI in education. Furthermore, the introduction of novel evaluation metrics such as transparency score, explainability ratio, and interpretability ratio alongside sparsity and sensitivity analyses, will contribute to a more robust understanding of XAI's role in educational settings. This comprehensive approach aims to bridge the gap between AI-driven predictions and real-world educational applicability, fostering a more equitable and comprehensible AI-assisted learning environment.

1.3 The Role and Challenges of Grade Prediction in the English Education System

In the UK, predicted grades play a pivotal role in student progression, particularly in determining post-16 educational and university pathways. However, the COVID-19 pandemic in 2020 exposed significant vulnerabilities in existing assessment systems. With the cancellation of formal GCSE and A-level examinations, Ofqual introduced the Direct Centre Performance model, an algorithm-driven framework intended to standardize grading in the absence of exams (Ofqual, 2020). The model, however, sparked national controversy when approximately 40% of students were downgraded compared to teacher predictions, disproportionately affecting those from state-funded schools (BBC News, 2020). Public outcry and parliamentary scrutiny led to the algorithm's abandonment, with final grades being based on teacher assessments. This event underscored the limitations of both traditional and algorithmic grading systems, reinforcing the need for more transparent, equitable, and trustworthy assessment approaches (Denes, 2023).

While teacher-predicted grades (TPGs) have long been central to the English education system, particularly for university admissions, their reliability and fairness have come under increasing scrutiny. Multiple studies highlight systemic issues in TPG accuracy, with significant instances of both over- and under-prediction. These discrepancies not only affect students' admission outcomes but can also undermine confidence and equity (McManus et al., 2021). Research indicates that students from lower socio-economic backgrounds are more likely to receive lower predicted grades despite similar academic capabilities (Stopforth, Gayle & Boeren, 2020). Biases based on ethnicity and gender have also been identified. Students from ethnic minority backgrounds and female students in STEM subjects are often under-predicted relative to their actual performance (Morris et al., 2021; Kim, Lee & Cho, 2022). These concerns highlight the urgent need to explore data-driven, explainable alternatives that ensure both accuracy and fairness.

General Certificate of Secondary Education (GCSE) exams, typically taken at age 16, serve as key milestones in England’s education system, determining eligibility for post-16 qualifications and influencing long-term educational trajectories. The 2017 transition from letter-based (A*–G) to numeric (9–1) grading aimed to provide finer differentiation of student achievement (Ofqual, 2020). However, the process of arriving at these grades whether through teacher predictions or algorithmic models continues to face scrutiny around fairness, transparency, and interpretability.

Machine learning (ML) has demonstrated potential in enhancing the accuracy of grade predictions by learning patterns from large volumes of student data. However, predictive accuracy alone is insufficient. Without transparency, even accurate models risk rejection by stakeholders due to perceived opaqueness and lack of accountability. XAI techniques such as

SHAP (SHapley Additive Explanations), LIME (Local Interpretable Model-Agnostic Explanations), ALE (Accumulated Local Effects), and PDP (Partial Dependence Plots) offer promising solutions to bridge this gap. These tools can help educators, students, and policymakers understand how predictions are generated and assess whether these insights align with pedagogical fairness and expectations (Adnan et al., 2022).

Emerging research is beginning to explore these approaches in educational contexts. For instance, Anders et al. (2020) applied ML techniques to predict student grades, but their models only correctly predicted around 25% of cases, with significant misclassification across others. Denes (2023) evaluated AI-based GCSE predictions in a selective independent school, achieving around 75% overall accuracy, with most predictions falling within one grade of the true outcome. However, notable outliers remained, with errors exceeding two or more grades thereby raising important concerns about the fairness and reliability of such tools in high-stakes assessments.

1.3.1 Current Approaches and Limitations in Grade Prediction.

Existing approaches to grade prediction in the UK predominantly rely on teacher assessments or statistical estimations, both of which are susceptible to bias and inconsistency. Teacher-predicted grades (TPGs), while grounded in professional judgement, are influenced by subjective expectations and implicit biases, particularly around student background, ethnicity, and gender (Murphy and Wyness, 2020; Magowan, 2023). These biases can unfairly impact students' academic trajectories, with under-prediction potentially limiting opportunities for further education or employment. While moderation procedures exist, they do not fully address these systemic disparities.

Machine learning offers a scalable and data-informed alternative. By leveraging historical performance data such as assessments, attendance, behaviour, and engagement, ML models can predict future outcomes with greater consistency. However, these models often function as "black boxes," lacking transparency in how predictions are made. This was evident during the 2020 grading controversy, where opaque algorithmic decisions prompted widespread distrust and protest (Ofqual, 2020). Educators and students alike expressed frustration over not understanding the rationale behind algorithmic outputs, highlighting the critical importance of model interpretability and stakeholder engagement.

Furthermore, predictive models trained on biased data risk perpetuating historical inequalities. If models are not corrected for embedded socio-economic, ethnic, or gender-related disparities, they may reinforce rather than mitigate disadvantage (Mehrabi et al., 2021). Biases in data collection, feature selection, and modelling processes can subtly influence predictions in ways that are difficult to detect without explainability mechanisms.

XAI provides a path forward by illuminating the inner workings of complex models. By offering intuitive, stakeholder-specific explanations, XAI facilitates trust and enables meaningful interpretation of predictions. For example, SHAP and LIME can show which features most influenced an individual student's grade prediction, while PDP and ALE can illustrate average effects across the dataset. These tools enable teachers and school leaders to interrogate model outputs, identify potential biases, and make more informed decisions about interventions and support.

Crucially, the success of AI in education depends not only on its technical performance but on its ability to complement human judgement. Stakeholder-centric explanations tailored to the information needs of students, teachers, and administrators can promote transparency, build confidence, and enhance the legitimacy of AI-assisted assessments (Fazil et al., 2024).

In summary, while current grade prediction approaches in the UK face multiple challenges including subjectivity, opacity, and bias, emerging XAI techniques offer the potential to address these limitations. This research builds on that promise by developing and evaluating a robust, interpretable, and fair AI-based grade prediction system tailored to the UK secondary education context.

1.4 Research Aim and Objectives

1.4.1 Research Aim

To adapt XAI techniques that enhance the transparency, interpretability, and trustworthiness of machine learning models used for predicting student performance in GCSE Mathematics and English within UK secondary education, thereby supporting informed and equitable educational decision-making.

1.4.2 Research Objectives

1. To design and implement a multi-target ensemble classifier and a multi-layer perceptron classifier for predicting GCSE Mathematics and English grades, using student performance data from a UK secondary school
2. To compare and evaluate the predictive performance of these models across key demographic and educational subgroups, such as socio-economic status, gender, and language background, ensuring fair generalisation across diverse student populations.
3. To aim for at least a 5% improvement in prediction accuracy over baseline models, where feasible, by optimizing feature selection, hyperparameter tuning, and data augmentation techniques tailored to GCSE performance prediction.
4. To adapt and apply XAI techniques such as SHAP, LIME, PDP, and ALE to the education domain, integrating domain-specific constraints (e.g., grading policies, assessment criteria, and learning behavior patterns) for improved contextual relevance.
5. To generate tailored explanations using XAI techniques that address the interpretability needs of multiple stakeholders including students, teachers, school administrators, and policymakers ensuring that AI-generated insights are actionable and pedagogically meaningful.
6. To develop and validate novel evaluation metrics for explainability, transparency, and interpretability such as fidelity**,** completeness**,** consistency**,** actionability**,** explainability ratio, and transparency score to assess the quality and trustworthiness of AI-generated predictions in educational contexts.

1.4.3 Research Questions

To address the overarching research aim, this study seeks to answer the following key research question:

How can explainable AI techniques be designed and adapted to improve the transparency, interpretability, and trustworthiness of machine learning models used in student performance prediction?

This is further broken down into the following sub-questions:

1. What are the key features and variables that most significantly influence student performance, and how do they vary across different educational contexts?

2. Which student performance indicators (e.g., exam scores, attendance, engagement levels) can be effectively predicted using machine learning models, and how can these predictions support educational decision-making?
3. How can complex machine learning models for student performance prediction be made more interpretable and transparent using XAI techniques such as SHAP, LIME, PDP, and ALE?
4. How can XAI-driven explanations be tailored to meet the needs of different stakeholders, including students, teachers, school administrators, and policymakers, ensuring AI-generated insights are interpretable and actionable?
5. What quantifiable evaluation metrics (e.g., explainability ratio, transparency score, and interpretability ratio) can be developed to assess the fidelity, completeness, consistency, actionability and trustworthiness of AI-driven student performance predictions?

1.5 Key Terms and Definitions

This section defines the foundational terms used throughout the thesis, grouped into three categories: (1) Artificial Intelligence and Machine Learning**,** (2) Explainable AI (XAI) Concepts**,** and (3) Educational Context and Stakeholder Engagement**.**

1.5.1 Artificial Intelligence and Machine Learning

1. **Artificial Intelligence (AI):** The field of computer science that enables machines to perform tasks typically requiring human intelligence, such as reasoning, learning, perception, and decision-making.
2. **Machine Learning (ML):** A subfield of AI focused on developing algorithms that learn from data to make predictions or decisions without explicit programming.
3. **Machine Learning Pipeline:** A structured end-to-end workflow that automates the development, training, validation, and deployment of machine learning models, enhancing reproducibility and efficiency.
4. **Ensemble Learning:** A machine learning approach that combines multiple models (e.g., decision trees, neural networks) to improve prediction accuracy and robustness.
5. **Multi-Layer Perceptron (MLP) Classifier:** A type of deep neural network used for classification tasks, consisting of multiple layers of interconnected neurons.
6. **Multi-Target Classification:** A machine learning technique that enables simultaneous prediction of multiple outcome variables, such as both English and Mathematics GCSE grades.

7. **Feature Selection:** The process of identifying and selecting the most relevant input variables from a dataset to enhance model accuracy and reduce complexity.
8. **Hyperparameter Tuning:** The process of optimizing configuration settings that govern a model's learning process (e.g., learning rate, tree depth) to improve performance.
9. **Data Augmentation:** A technique used to artificially expand the training dataset to improve generalizability and reduce overfitting.
10. **Synthetic Minority Over-sampling Technique (SMOTE):** A method for addressing class imbalance by generating synthetic examples of minority class instances, enhancing fairness in classification tasks.

1.5.2 Explainable AI (XAI) Concepts

11. **Explainable Artificial Intelligence (XAI):** A branch of AI focused on making machine learning models transparent and interpretable by providing human-understandable explanations for predictions.
12. **Transparency in AI:** The degree to which the internal mechanisms of an AI model are visible and understandable—such as model architecture, parameters, or training processes.
13. **Interpretability in AI:** The extent to which a human, particularly a non-technical user, can comprehend and derive meaning from an AI model's output.
14. **Trustworthiness in AI:** The extent to which AI-generated predictions are reliable, unbiased, and ethically aligned with human values, especially in high-stakes applications like education.
15. **Bias in AI Models:** Systematic errors that unfairly disadvantage certain groups, often arising from unbalanced or unrepresentative training data or model assumptions.
16. **Fairness in AI:** The commitment to ensuring AI systems do not discriminate against specific demographic or socio-economic groups and promote equity in decision-making.
17. **Shapley Additive Explanations (SHAP):** A model-agnostic technique that attributes a prediction to individual features based on their marginal contribution.
18. **Local Interpretable Model-Agnostic Explanations (LIME):** A method that explains individual predictions by approximating the original model with a simpler, local interpretable model.

19. **Partial Dependence Plots (PDP):** Visual tools that depict the average effect of one or more input variables on model predictions.
20. **Accumulated Local Effects (ALE):** An alternative to PDPs that accounts for feature interactions and avoids extrapolation, offering more accurate interpretation in complex models.
21. **Evaluation Metrics in XAI:** Quantitative measures used to assess the effectiveness of explanations provided by AI models, especially in terms of their clarity and utility. These include:

- **21.1 Fidelity:** How well the explanation reflects the true reasoning of the model.
- **21.2 Completeness:** Whether the explanation includes all critical factors influencing the outcome.
- **21.3 Consistency:** The stability of explanations across similar cases.
- **21.4 Actionability:** Whether the explanation leads to decisions or interventions that improve outcomes.

1.5.3 Educational Context and Stakeholder Engagement

22. **Student Performance Prediction:** The use of ML techniques to forecast academic performance based on prior data such as attendance, grades, behavior, and engagement metrics.
23. **Educational Stakeholders:** Individuals or groups impacted by AI use in education, including students, teachers, school leaders, policymakers, and parents.

1.6 Research Methodology

This thesis employs a multi-step approach to adapting XAI techniques for student performance prediction:

1. Literature Review: A comprehensive analysis of the role of AI in education, explores machine learning models for prediction and a comprehensive analysis of existing XAI methods, focusing on their applicability to educational data and the specific needs of educational stakeholders.
2. Data: Thorough presentation of data selection, data pre-processing and extensive exploratory data analysis.

3. Model Selection: Development and evaluation of various machine learning models, including random forests, gradient boosting, and deep learning models, trained on student performance datasets.
4. XAI Adaptation
5. User-Centric Evaluation: Testing the XAI methods with real-world educational stakeholders through user studies, focusing on how understandable and actionable the model explanations are.
6. Metric Development: Creation of new metrics to evaluate the transparency, interpretability, and actionability of XAI models, with particular focus on balancing accuracy with explainability.
7. Research Conclusion: Summary of findings, research limitation and challenges, summary of findings, key contribution and future directions.

1.7 Thesis Structure

This thesis is organized into seven chapters, each building on the preceding one to present a comprehensive investigation into explainable machine learning for predicting student performance in secondary education in the UK.

Chapter 1 introduces the research context, objectives, and guiding questions, and outlines the significance and scope of the study.

Chapter 2 presents a critical review of existing literature on student performance prediction, machine learning in education, algorithmic bias, and explainable AI (XAI). It identifies key research gaps, especially in the application of XAI techniques to the UK GCSE context.

Chapter 3 details the research methodology, including data collection and preparation, feature selection, model development, and the design of stakeholder-focused evaluation studies. It also introduces the explainability metrics and evaluation techniques used in this study.

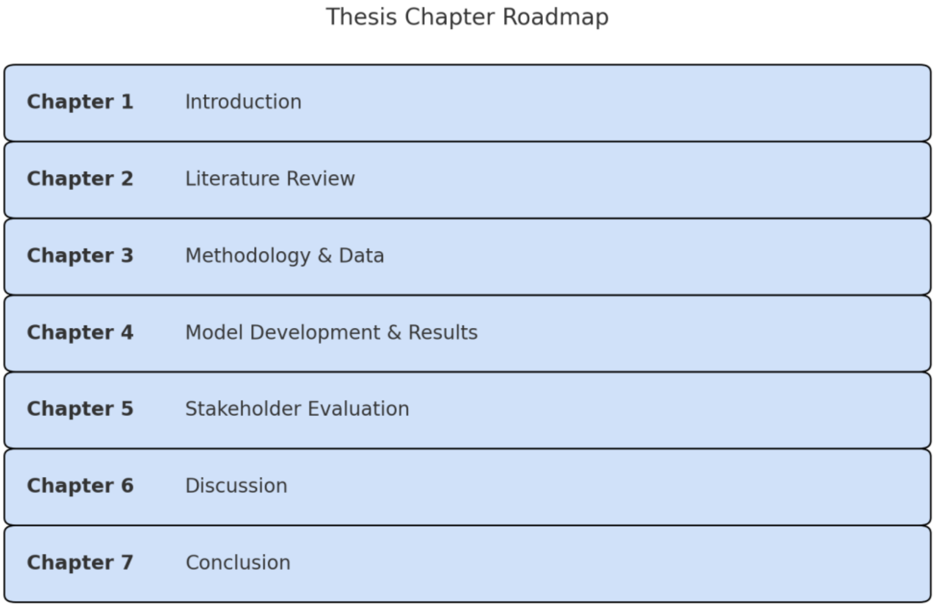
Chapter 4 provides an in-depth account of model implementation and performance. It compares predictive models using a range of performance and explainability metrics and analyses the trade-offs between accuracy and transparency.

Chapter 5 presents the results of stakeholder evaluations, including feedback from teachers and policymakers on the interpretability and usefulness of the model outputs. It also compares decision-making outcomes with and without access to explainability insights.

Chapter 6 discusses the findings in relation to existing research, highlighting both theoretical and practical implications. It reflects on the ethical dimensions of AI in education and the challenges of implementing fair and interpretable models in real-world settings.

Chapter 7 concludes the thesis by summarizing key contributions, acknowledging limitations, and suggesting directions for future research, particularly the development of human-centered, context-sensitive XAI solutions in education.

Figure 1 presents a visual overview of the thesis structure which begins with foundational context (Chapter 1), moves through literature and methodology (Chapters 2–3), and culminates in findings, stakeholder validation, and final conclusions (Chapters 4–7).



Thesis Chapter Roadmap

Chapter 1	Introduction
Chapter 2	Literature Review
Chapter 3	Methodology & Data
Chapter 4	Model Development & Results
Chapter 5	Stakeholder Evaluation
Chapter 6	Discussion
Chapter 7	Conclusion

Figure 1: Thesis Chapter Roadmap

1.8 Chapter Conclusion

This chapter has outlined the background, challenges, and objectives of this research on XAI in student performance prediction. The study highlights the need for transparency, fairness, and accuracy in AI-driven educational models while addressing bias and ethical concerns**.** By integrating XAI techniques, this research aims to bridge the gap between AI's predictive power and the need for human-interpretable decisions. The subsequent chapters build upon this foundation, exploring relevant literature, methodological design, model development, and stakeholder evaluations. Through this progression, the thesis seeks to contribute to a fairer, more transparent approach to AI-supported assessment in UK secondary education.

Chapter 2 – Literature Review

Predicting student performance is a complex task influenced by multiple variables, including demographic, socio-economic, psychological, and environmental factors. These factors shape students' actions, decision-making processes, and learning outcomes (Hayes, 2021). Demographic parameters, often employed in population studies, provide statistical insights into student characteristics such as employment status, education level, income, and socio-economic background (Basu & Goldhaber-Fiebert, 2015). Understanding these variables is crucial for educators, researchers, and policymakers to address gaps in education and design targeted interventions that support student success (Hayes, 2021).

Machine learning (ML) has become a powerful tool for predicting student performance by analysing these variables and offering data-driven insights. ML models support early interventions, personalised learning, and curriculum adjustments based on predictive patterns (Ayienda et al., 2021; Tong & Li, 2024; Burke et al., 2024). However, the application of ML in educational settings presents challenges related to bias, fairness, and interpretability. Ensuring that these models provide accurate and meaningful predictions requires careful selection of input variables and an understanding of their implications for student learning.

2.1 Literature Review Strategy

This literature review adopted a thematic synthesis approach to systematically examine the current body of research on student performance prediction using ML and XAI. The review aimed to identify factors influencing academic outcomes, assess the effectiveness of predictive models, and evaluate the adoption of XAI techniques in educational contexts.

2.1.1 Search Databases and Keywords

The literature search was conducted using academic databases including:

- Scopus
- Web of Science
- IEEE Xplore
- Google Scholar
- ERIC

Keywords and search strings included:

- "Student performance prediction"
- "Machine learning in education"
- "Predicting GCSE outcomes"
- "Educational fairness and AI"
- "Explainable artificial intelligence in education"
- "XAI for student success prediction"

2.1.2 Inclusion and Exclusion Criteria

- Inclusion: Peer-reviewed journal articles, conference papers, and systematic reviews published between 2015 and 2024, focusing on educational AI, student performance prediction, or XAI applications.
- Exclusion: Non-English papers, studies with inadequate methodology transparency, and those not involving predictive analytics.

2.1.3 Review Process

Over 120 studies were identified. After title and abstract screening, 67 studies were shortlisted for full review. The studies were coded thematically across six core domains:

1. Demographic and Socioeconomic Factors
2. Behavioral and Psychological Traits
3. Digital Literacy and Technology Use
4. Feature Engineering and Preprocessing
5. ML Models and Hybrid Techniques
6. Fairness, XAI, and Ethical Considerations

This process enabled the identification of key gaps in the literature, particularly around fairness-aware modelling, explainability in high-stakes environments like GCSEs, and stakeholder-specific interpretability.

2.2 Demographic and Socioeconomic Factors in Academic Performance

Academic performance is shaped by a complex interplay of demographic, socio-economic, psychological, and environmental factors. Understanding these variables is crucial for developing accurate predictive models using machine learning (Suleiman, 2023; Saha et al., 2022; Munir et al., 2023). This section critically examines key demographic factors affecting student performance, drawing on recent empirical studies and highlighting their implications for educational analytics and predictive modelling.

2.2.1 Socioeconomic Status (SES) and Academic Performance

Numerous studies have established a strong link between socio-economic status (SES) and student achievement. Sammons (1995) analysed educational disparities over nine years, emphasising how factors such as family income, parental education, and neighbourhood characteristics shape academic trajectories. Similarly, Nawa et al. (2020) found that medical students from non-urban areas in Japan were 7.2 times more likely to experience academic failure, underscoring how geographic disparities impact student success.

In South Africa, Luwes et al. (2017) discovered that students whose native language was Afrikaans or Zulu outperformed those from different linguistic backgrounds. This finding aligns with Masud et al. (2019), who demonstrated that socio-economic conditions, parental education, and parenting styles strongly predicted student performance in Pakistan. Additionally, Wickrama et al. (2021) argued that early socio-economic disadvantages not only affect immediate academic success but also perpetuate long-term economic hardship, reinforcing cycles of inequality. Similarly, Jasim (2020) confirmed that family social status impacts academic achievement, suggesting that peer influences also play a crucial role in reinforcing these outcomes.

In contrast, Al-Azawei & Al-Masoudy (2020) argue that while demographic characteristics do influence learning outcomes, behavioural and engagement factors often overshadow these effects. Their findings suggest that students' proactive behaviours significantly enhance academic performance, regardless of their socio-economic status. These discrepancies indicate that ML models predicting student performance must carefully weigh the relative importance of SES against behavioural traits to improve accuracy and fairness.

2.2.2 Gender and Academic Achievement

The role of gender in academic achievement has long been debated, with researchers exploring differences in performance between male and female students. Some studies suggest that gender plays a key role in shaping academic success, while others argue that it is only one of many contributing factors. Malik et al. (2024) found that gender influences academic performance, particularly through variations in academic anxiety and motivation. Their research indicates that female students often report higher levels of academic anxiety, which can negatively impact their performance. However, Khan et al. (2024) challenge this perspective, suggesting that when self-efficacy and learning strategies are considered, gender differences become statistically insignificant. This implies that disparities in academic success may be driven more by psychological and engagement factors than inherent cognitive differences.

Beyond individual academic capabilities, social and cultural expectations also shape how students perceive their abilities and cope with challenges. Malik et al. (2024) emphasise that gender norms influence self-perception and coping strategies, which, in turn, affect academic trajectories. Yao et al. (2024) support this claim, noting that while female students tend to report higher self-esteem, this advantage does not always translate into superior academic outcomes. These findings suggest that the relationship between gender, motivation, and achievement is complex and cannot be explained by gender alone. Instead, psychological and social influences play a crucial role in shaping students' academic experiences.

Despite evidence suggesting gender-based differences in academic performance, other studies have shown that once additional factors are considered, gender’s influence diminishes. Khan (2021) examined how personality traits interact with academic success and found that female students benefited from openness to experience, while male students performed better when emotional stability was a contributing factor. This highlights the importance of looking beyond gender and considering personality and behavioural traits in analysing academic achievement.

2.2.3 Parental and Cultural Influences on Learning

Parental involvement is widely recognised as a key determinant of student success, though its impact varies based on parenting styles, socio-economic background, and student personality. Research suggests that authoritative parenting, which balances encouragement with structure, fosters self-efficacy and improves academic outcomes (Hayek et al., 2022; Nair et al., 2024). Similarly, An et al. (2019) found that parental engagement in STEM subjects positively influences academic performance.

However, the effectiveness of parental involvement depends on the child's individual needs. Saka (2022) observed that some students thrive under structured guidance, while others benefit more from independent learning. Mihret et al. (2019) reported that in certain cultures, authoritarian parenting demanding high academic performance without emotional support can yield high grades, whereas other studies suggest that such rigidity can increase stress and disengagement (Ma et al., 2021; Zahedani et al., 2016).

Socio-economic factors also influence parental involvement. Financial constraints and work commitments may limit parents' ability to engage in their child's education (Mugumya et al., 2023). However, Butler (2021) challenges the assumption that lower-income parents are less invested in education, highlighting their strong commitment despite resource limitations. These nuances suggest that ML models must consider a wide range of parental engagement indicators beyond income or education level to provide more accurate predictions (Bettencourt et al., 2020).

2.2.4 Psychological and Behavioral Factors

Psychological attributes such as self-esteem and self-regulation play a significant role in academic success. Vacalares et al. (2023) found a strong correlation between self-esteem and academic performance, though other studies suggest that academic success itself reinforces self-esteem (Nguyen et al., 2019). Additionally, students with low self-esteem are more likely to experience anxiety and stress, negatively impacting their ability to concentrate (Nguyen et al., 2019).

Self-regulation, or the ability to manage emotions, focus, and sustain effort, is another key predictor of academic achievement. Research shows that students with poor self-regulation are less engaged and achieve lower academic outcomes (Atasoy & Pekel, 2021). Those with strong self-regulation skills, however, are more likely to set realistic goals and persist through challenges (Roth et al., 2017).

2.2.5 Technology and Digital Literacy

In today's educational landscape, digital literacy has become a key factor in student success. As schools and universities increasingly integrate technology into their teaching methods, understanding how digital skills impact academic performance has gained growing importance. While digital competency enables students to engage more effectively with learning materials, it is also shaped by factors such as socioeconomic background, cultural context, and access to technology.

Research consistently shows a strong correlation between digital competencies and academic achievement. Mehrvarz et al. (2021) found that students with advanced digital skills tend to perform better in their studies, a finding supported by Zhao et al. (2021), who highlighted digital informal learning as a significant factor in educational success. Their analysis suggests that students who are proficient with technology are better equipped to navigate academic challenges and engage with course content. However, some researchers caution against viewing digital literacy as the sole determinant of academic performance. Weli et al. (2024) argue that while technology enhances learning, it does not automatically bridge academic gaps for at-risk students. Instead, they emphasize the importance of early identification of struggling learners and the use of educational technology to provide targeted interventions.

Despite the advantages of digital literacy, not all students have equal access to technology, leading to disparities in academic performance. This persistent digital divide disproportionately affects students from lower-income backgrounds who may lack reliable internet access, personal devices, or exposure to digital tools at home. Weli et al. (2024) highlight that, factors such as parental education and socioeconomic status significantly influence a student's ability to develop digital competencies. Zhao et al. (2021) echo this concern, noting that digital proficiency is often linked to a student's financial background. Without equal access to technology, students from disadvantaged backgrounds face an additional barrier to academic success.

Lucas et al. (2022) raise another issue, arguing that self-reported digital competency may not always reflect actual proficiency. Many students perceive themselves as tech-savvy but may lack the practical skills necessary for effective digital learning. This gap between perceived and actual competence suggests that educators and researchers should adopt more rigorous assessment methods to understand students' digital abilities more accurately.

Beyond socioeconomic disparities, cultural factors also influence digital literacy. Morales et al. (2024) found that students' cultural backgrounds shape their attitudes toward technology and learning, affecting how they engage with digital tools. Some cultures emphasize traditional learning methods, while others encourage technology-driven education. These differences suggest that educational institutions must adopt culturally responsive teaching methods to ensure that digital learning strategies are effective across diverse student populations.

Given the complex relationship between digital literacy and academic performance, machine learning models designed to predict student success must incorporate multiple variables. First, predictive models should integrate not only digital competency levels but also factors such as access to technology, socioeconomic background, and cultural influences. By accounting for these elements, machine learning can provide a more holistic understanding of student performance (Weli et al., 2024; Zhao et al., 2021). Second, advanced machine learning techniques can be used to model non-linear relationships between digital literacy and academic success, uncovering patterns that simpler models might overlook (Yahaya & Ogundola, 2024). Third, predictive models must be adaptable to different educational and cultural contexts. Kwiatkowska & Wiśniewska-Nogaj (2022) argue that localized data collection is essential for ensuring that models accurately reflect the realities of different student populations.

Another critical consideration is addressing the digital divide in predictive analytics. Machine learning models should be designed to identify students who are at risk due to limited digital access and provide recommendations for interventions. Weli et al. (2024) and Lucas et al. (2022) emphasize that predictive models should help schools and policymakers allocate resources effectively to ensure that all students have the opportunity to develop strong digital skills. Additionally, continuous assessment is crucial. Rather than relying on a one-time measurement of digital literacy, ongoing tracking of students' technology use and skills development can improve the accuracy of machine learning predictions (Mehrvarz et al., 2021; Litiņa & Svētiņa, 2023).

The existing literature highlights digital literacy as an important factor in student success while also revealing the complexities surrounding access, cultural attitudes, and the digital divide. As educational institutions continue to adopt technology-driven learning methods, machine learning models must incorporate a diverse range of data to ensure accurate and fair predictions. By considering socioeconomic disparities, cultural influences, and evolving digital skills, these models can help create more equitable learning environments and better support students in achieving academic success.

2.2.6 Implications for Machine Learning in Student Performance Prediction

The reviewed literature underscores the complexity of student performance, demonstrating that it cannot be accurately predicted using isolated demographic variables. While factors such as SES, gender, and parental involvement provide valuable insights, they do not operate in isolation. Behavioural, psychological, and technological dimensions must also be considered to develop a holistic understanding of academic achievement.

Machine learning has become a valuable tool for predicting student performance, but it also presents challenges. One of the biggest concerns is bias in training data. If predictive models are trained on datasets that reflect existing inequalities, they risk reinforcing rather than correcting them (Imran et al., 2020). Ensuring that training datasets are diverse and representative is essential for producing fair and accurate predictions. Additionally, predictive models should not merely identify performance trends but also provide actionable insights that educators can use to support students.

Machine learning has the potential to transform education by enabling early interventions and personalised learning strategies. However, for these models to be truly effective, they must be designed to reflect the complexity of student performance. By incorporating behavioural data, cultural influences, and psychological attributes, predictive models can move beyond simple classifications and contribute to more meaningful and equitable educational interventions. This research will therefore focus on refining feature selection methods and ensuring that predictive models remain interpretable and actively mitigate bias.

2.3 Feature Engineering in Educational Performance Prediction

Feature engineering is a foundational process in machine learning that involves transforming raw data into structured, informative inputs that improve the performance of predictive models. In the domain of educational data mining, feature engineering plays a particularly vital role in capturing the multifaceted nature of student learning and behavior. By extracting and refining features such as prior academic achievement, attendance records, engagement patterns, and socio-demographic variables, researchers can develop models that more accurately reflect students’ academic trajectories.

Jiang et al. (2022) define feature engineering as a process of understanding the application context and the intrinsic characteristics of the data. In the context of education, this involves recognizing the behavioral, academic, and contextual variables that influence student success. Effective feature engineering can bridge the gap between raw educational data and meaningful insights, enabling predictive models to uncover latent patterns that may otherwise remain hidden.

Research in the field has consistently identified certain features as particularly predictive of student performance. Attendance and prior grades are among the most reliable indicators, with multiple studies emphasizing their centrality in forecasting academic outcomes. Mustapha (2023) underscores the importance of integrating a diverse set of features, including not only academic metrics but also engagement indicators and behavioral data, to capture a more nuanced profile of student performance. Socio-demographic variables such as gender, parental education, and socio-economic background have also been found to correlate strongly with learning outcomes, particularly in their influence on access to educational resources, support structures, and overall learning opportunities.

In addition to traditional academic records, modern educational systems increasingly incorporate data from digital learning platforms. Learning Management Systems (LMS) and Massive Open Online Courses (MOOCs) generate a wealth of behavioral data, such as clickstream logs and interaction timestamps. Kőrösi and Farkas (2020) demonstrated how features extracted from raw clickstream data, including frequency of platform use and time spent on tasks, can be used to predict student success in online learning environments. Such data enables a shift from static to dynamic models of student performance prediction, allowing for more responsive and real-time analytics.

To effectively handle diverse and often complex educational data, a range of feature engineering techniques has been applied. Categorical variables such as course enrolments or demographic characteristics are often encoded using methods like one-hot encoding or ordinal encoding, which make the data compatible with most machine learning algorithms. Feature transformation techniques such as scaling, normalization, and binning are also commonly employed to ensure that continuous variables contribute meaningfully to model learning. Mustapha (2023) highlights the importance of combining multiple engineering strategies to enhance the representational value of student performance data.

Dimensionality reduction techniques further contribute to the refinement of feature sets. Principal Component Analysis (PCA), for example, has been employed to reduce data complexity while preserving key information. These techniques not only improve computational efficiency but also enhance model accuracy by removing redundant or irrelevant features. Feature selection methods, such as recursive feature elimination and information gain, allow researchers to identify the most informative features while reducing the risk of overfitting. Naseer et al. (2020) found that models using optimized feature sets generated through information gain significantly outperformed those using raw data inputs.

Beyond predictive performance, feature engineering also plays a key role in ensuring model fairness and interpretability. Al-Ahmad et al. (2022) emphasize the ethical implications of feature selection, particularly when dealing with sensitive attributes such as socio-economic status or ethnicity. Poorly selected features can unintentionally introduce bias into prediction models, disadvantaging certain student groups. Thoughtful feature engineering, combined with fairness-aware modelling practices, can help mitigate such risks by promoting equitable evaluation criteria. Moreover, by identifying and retaining only the most influential features, models become more interpretable, aiding educators and decision-makers in understanding the rationale behind predictions.

In sum, feature engineering is an essential step in the development of robust, fair, and interpretable predictive models in education. The literature highlights a broad consensus on the importance of features such as attendance, prior achievement, and socio-demographics, while also acknowledging the growing relevance of behavioral and engagement data from digital learning platforms. A range of techniques, from encoding to dimensionality reduction, has been effectively employed to refine feature sets and improve model performance. Importantly, feature engineering also supports fairness and transparency, ensuring that predictive models not only achieve high accuracy but also uphold ethical standards in educational assessment.

2.4 Machine Learning Models for Student Performance Prediction

ML and AI have found extensive application across various domains, including education, advertising, financial analysis, and fraud detection (Hutt et al., 2019). In the educational sector, ML models are increasingly utilized to facilitate data-driven decision-making by analysing diverse student attributes to predict academic outcomes. However, challenges such as bias in training data, model interpretability, and fairness persist, raising important concerns regarding the effectiveness and ethical implications of ML applications in educational settings (Hutt et al., 2019). This review explores different ML techniques used for student performance prediction, compares their effectiveness, and highlights ensemble learning and multilayer perceptron (MLP) as key methodologies for achieving optimal results. This literature review adopts a thematic approach, critically analysing existing studies on machine learning models in education, comparing different predictive techniques, and evaluating the role of XAI in enhancing the reliability of these models.

2.4.1 Advances in ML for Predicting Educational Outcomes

The ability of ML algorithms to analyse large-scale datasets has led to significant advancements in academic performance prediction. Alyahyan and Düştegör (2020) highlighted the potential of these methods to uncover patterns in student behaviour and academic trends. Their research demonstrated that ML models can provide more accurate predictions compared to traditional statistical approaches, enabling educators to identify at-risk students early.

Both supervised and unsupervised learning techniques have been employed for performance prediction. While supervised learning methods such as Decision Trees (DTs) and Artificial Neural Networks (ANNs) have been widely used, unsupervised learning has been leveraged to uncover latent patterns in student data (Bajari et al., 2015). However, the effectiveness of these models varies depending on dataset characteristics and the inclusion of key predictive features.

2.4.2 Performance of ML Techniques: Supervised Learning Approaches

A wide range of studies have explored the effectiveness of various machine learning (ML) algorithms in predicting student performance, highlighting both their predictive capabilities and associated limitations. Wu (2021) employed Decision Trees (DTs) and Artificial Neural Networks (ANNs) to classify students based on multiple input parameters, showing the utility of these models in educational analytics. Decision Trees, in particular, have been favoured for their transparency and ease of interpretation. For instance, Rizvi, Rienties, and Khoja (2019) used a DT model to examine how demographic factors such as socioeconomic status, gender, and disabilities influence student outcomes. However, DTs are known to overfit training data, reducing their generalisability across different contexts.

In contrast, ANNs have become increasingly popular due to their ability to capture complex, non-linear relationships in educational data. Kehinde et al. (2021) developed an ANN model using prior academic records and demographic features to predict student admissions, achieving a precision rate of 92.3%. Their results demonstrated the strength of deep learning in modelling educational outcomes, though the study also acknowledged the challenge of limited interpretability, which can be problematic for educators seeking clear, actionable insights.

Cruz-Jesus et al. (2020) conducted a comprehensive comparison of several ML techniques such as ANNs, Decision Trees, Extra Trees, Random Forests (RF), Support Vector Machines (SVM), K-Nearest Neighbours (KNN), and Logistic Regression. They found that ensemble methods such as Extra Trees and RF consistently outperformed individual classifiers, achieving accuracy rates around 75%. Ensemble models like RF, which aggregate predictions from multiple DTs, were particularly effective in reducing overfitting and improving robustness. This was corroborated by Baashar et al. (2021), who also found that RF and ANNs achieved the highest predictive performance when applied to educational datasets.

Support Vector Machines have also shown strong classification performance, especially with high-dimensional data (Su et al., 2021). However, their computational intensity and sensitivity to parameter tuning can limit their practical application in large-scale, real-time educational settings. KNN models, though simple and interpretable, are susceptible to noise and rely heavily on distance metrics, which can impact performance when dealing with diverse and imbalanced datasets (Kavitha et al., 2022). As such, KNN is often considered more suitable as a baseline model rather than a primary predictive tool.

Naïve Bayes (NB), known for its simplicity and scalability, can effectively handle high-dimensional data. Nonetheless, its core assumption of feature independence frequently does not hold in educational contexts, where attributes such as attendance, socio-economic status, and prior achievement are often interrelated (Winzeck et al., 2018).

Alyahyan and Düştegör (2020) also compared DTs, SVMs, and ANNs, noting that while DTs offer interpretability, they are prone to overfitting. SVMs delivered strong results in complex feature spaces, though required extensive tuning, and ANNs particularly multilayer perceptron (MLP) achieved the highest accuracy, albeit at the cost of transparency.

While high-performing models such as ANNs and ensemble techniques demonstrate promising results in terms of predictive accuracy, several studies have emphasised the need for explainable artificial intelligence (XAI). For example, both Baashar et al. (2021) and Cruz-Jesus et al. (2020) stressed that despite the superior accuracy of models like RF and Extra Trees, their lack of transparency can hinder adoption in education, where trust and accountability are critical.

2.4.3 Hybrid Models and Ensemble Techniques

The application of hybrid classifiers in predicting student performance has garnered substantial interest within the educational data mining (EDM) community. These classifiers, which integrate multiple machine learning (ML) algorithms, are often found to outperform traditional single-model approaches by capitalizing on the strengths and compensating for the weaknesses of individual models. Ensemble techniques such as bagging, boosting, stacking, and voting fall under this umbrella and are increasingly recognized for their ability to enhance predictive performance in complex educational contexts.

A key advantage of hybrid classifiers lies in their ability to harness algorithmic diversity. Hsu (2017) emphasizes that such diversity among base learners is not merely beneficial but essential for improving prediction outcomes. The rationale is that different algorithms may capture distinct patterns or handle noise and bias differently, leading to more generalizable models. Evangelista and Sy (2022) empirically support this view by demonstrating that both homogeneous (e.g., bagging and boosting) and heterogeneous (e.g., stacking and voting) ensembles yield superior prediction accuracy compared to individual classifiers. Similarly, Francis and Babu (2019) extend this discussion by incorporating clustering into hybrid classification frameworks, underscoring that combining different data mining techniques can further refine predictive insight in academic settings.

Several studies provide empirical evidence of the effectiveness of hybrid approaches. Siddique et al. (2021) report a hybrid model combining multilayer perceptron (MLP), decision trees, and logistic regression achieved an accuracy of 98.5%, significantly outperforming standalone models. However, this impressive performance comes at the cost of increased computational complexity and reduced model interpretability. These factors may hinder real-world adoption in educational institutions. The same study reiterates these findings when evaluating MLP alongside J48 decision trees and bagging/boosting methods, with MLP emerging as the top performer. While accuracy gains are evident, there is limited discussion on how these models fare in terms of fairness, transparency, or usability by educators which are dimensions increasingly important in educational AI applications.

Ayienda et al. (2021) similarly found that hybrid classifiers combining KNN, SVM, Naïve Bayes, MLP, and linear regression yielded a predictive accuracy of 97.6%. Yet, despite the high performance, the study lacks granularity in explaining how each algorithm contributed to the ensemble or whether the model was evaluated across different student subgroups. This absence of fairness-aware evaluation raises questions about potential bias propagation in ensemble systems, a concern echoed in recent XAI and ethical AI literature.

Voting-based ensemble methods, particularly those utilizing majority or weighted voting, have also shown promise in improving robustness. Ostvar and Moghadam (2020) describe how aggregating classifier outputs through simple or weighted voting reduces variability and error, leading to more stable predictions. Haque et al. (2016) expand on this by integrating genetic algorithms to optimize ensemble composition, demonstrating the method's adaptability across classification tasks. While these innovations offer potential, they again introduce additional complexity, which could compromise transparency, an important consideration in high-stakes domains like education. As Kim et al. (2014) argue, hybrid voting mechanisms can refine outcomes further by weighting classifiers according to past performance, but such strategies may also obscure the decision-making process, making explanations harder to generate and understand.

Gajwani and Chakraborty (2020) reinforce the effectiveness of hybrid methods by demonstrating that combining random forests with other classifiers improves prediction accuracy in academic performance forecasting. However, like other studies, they focus predominantly on accuracy metrics, without adequately addressing practical concerns such as computational efficiency, scalability to large student datasets, or explainability for non-technical stakeholders like parents, teachers and school administrators.

Ashraf et al. (2020) conclude that hybrid classification techniques not only improve accuracy but also increase robustness, a desirable trait in dynamic and diverse educational environments. Nonetheless, across much of the reviewed literature, there remains a tendency to prioritize predictive performance over interpretability, fairness, and contextual applicability which are elements critical for ethical and effective AI deployment in education.

In summary, the literature consistently highlights the predictive superiority of hybrid classifiers in student performance prediction tasks. Ensemble techniques whether through bagging, boosting, stacking, or voting have proven effective in leveraging algorithmic diversity to enhance accuracy and robustness. However, a critical review reveals a recurring emphasis on accuracy at the expense of other vital dimensions such as interpretability, fairness, and real-world ‘deployability’. This work aims go beyond performance metrics to address these gaps, especially given the ethical implications of automated decision-making in education. Only by balancing predictive power with transparency and equity can hybrid classifiers be responsibly integrated into educational settings.

2.4.4 Role of MLP in Predictive Modelling

Among the various machine learning techniques applied in educational data mining, the Multilayer Perceptron (MLP), a type of feedforward artificial neural network, has consistently demonstrated exceptional predictive capabilities. MLP's architecture which comprises of multiple hidden layers and non-linear activation functions enables it to model both linear and complex non-linear relationships in data, making it particularly well-suited for predicting student academic outcomes. Its deep learning structure captures intricate interactions between features such as prior achievement, attendance, socio-economic status, behavioural patterns, and learning environments, all of which play crucial roles in student success.

Several studies provide empirical support for the use of MLPs in educational settings. Siddique et al. (2021), for instance, identified MLP as the highest-performing model in their comparative study, achieving a predictive accuracy of 98.5%, significantly outperforming other standalone classifiers. Similarly, Ghorbani and Ghousi (2020) demonstrated the effectiveness of MLP in early identification of students at risk of academic underperformance, thereby facilitating timely interventions. Ayienda et al. (2021) also incorporated MLP into a hybrid ensemble with other classifiers such as KNN, SVM, and Naïve Bayes, reporting a strong predictive accuracy of 97.6%. These studies collectively reinforce MLP's suitability for modelling the complexity inherent in educational datasets.

In addition to its accuracy, MLP offers high adaptability across diverse data types and educational contexts. Its capacity to process both numerical and categorical data, along with its scalability across different input-output formats, makes it a valuable model for a wide range of prediction tasks. This adaptability is particularly important in education, where data sources can vary considerably in structure and quality.

However, MLP's strengths are tempered by certain limitations. One of the most frequently cited challenges is the lack of interpretability. As with many artificial neural networks, MLP functions as a black-box model, offering limited insight into how predictions are generated. This characteristic poses a barrier in educational contexts, where stakeholders such as teachers, students, and administrators require transparent explanations to trust and act upon model outputs. Comparisons across studies have shown that simpler models, including logistic regression, decision trees, and random forests, often provide more interpretable outputs while maintaining reasonably good performance (Rizvi, Rienties, & Khoja, 2019; Cruz-Jesus et al., 2020). Although these models may not match MLP's predictive accuracy, their transparency offers better alignment with ethical and practical considerations in education.

MLPs have also been successfully deployed within hybrid and ensemble frameworks to boost performance and stability. When combined with ensemble strategies such as bagging, boosting, or weighted voting, MLPs contribute significantly to model robustness by enhancing classifier diversity. However, the computational complexity and training time of MLP-based models must also be considered, particularly in institutions with limited technical resources.

While the interpretability of MLP remains a challenge, this can be mitigated through the integration of explainable AI techniques and hybrid modelling approaches. Balancing performance with transparency is essential in educational contexts, and MLP offers a strong foundation upon which reliable, adaptable, and interpretable predictive systems can be built.

2.4.5 Comparative Analysis of Machine Learning Techniques for Student Performance Prediction

The increasing use of machine learning (ML) in educational data mining has enabled more accurate predictions of student academic performance, allowing institutions to implement proactive support strategies. The selection of an appropriate ML model is influenced not only by its predictive accuracy but also by its interpretability, scalability, and suitability for the specific dataset in use. As Baashar et al. (2021) highlight, the effectiveness of a model is context-dependent and varies based on dataset complexity, class imbalance, and feature dimensionality.

Artificial Neural Networks (ANNs), particularly the Multilayer Perceptron (MLP), have shown high accuracy in modelling non-linear relationships. Studies by Cruz-Jesus et al. (2020) and Ghorbani and Ghousi (2020) demonstrate MLP's capacity to learn intricate student-related patterns. However, the model's black-box nature and computational demands limit interpretability, which is essential in educational decision-making.

Decision Trees (DTs) are valued for their interpretability and rule-based structure (Baashar et al., 2021). Despite this, they tend to overfit noisy datasets and underperform in complex scenarios. Random Forests (RF) mitigate this issue by aggregating multiple trees, improving generalization and offering feature importance insights (Rodriguez-Hernandez et al., 2021). However, their interpretability still falls short in fine-grained educational contexts.

Support Vector Machines (SVMs) are powerful classifiers in high-dimensional spaces but are computationally intensive and lack transparency (Siddique et al., 2021). Similarly, K-Nearest Neighbors (KNN) performs well in small datasets but is sensitive to irrelevant features and scales poorly (Francis & Babu, 2019). Naïve Bayes (NB) models are scalable and effective for categorical data but assume feature independence, which is rarely valid in educational contexts (Sokkhey, 2020).

Ensemble methods such as gradient boosting, particularly XGBoost, have achieved state-of-the-art performance across educational benchmarks. These models iteratively improve accuracy and reduce overfitting. However, they are complex and computationally demanding. To address the need for efficient predictive modeling in educational performance, Histogram-Based Gradient Boosting (HGB) has emerged as a viable and effective variant that retains high performance while significantly reducing computational load. This method discretizes continuous features into histograms, facilitating faster training, better handling of missing values, and improved support for categorical variables. Importantly, HGB leverages the principles of gradient boosting, using histogram representations to enhance efficiency without sacrificing prediction accuracy (LIU, 2024; Setyarini et al., 2024).

Recent studies underscore the effectiveness of HGB in predicting academic success. Liu emphasizes that the Histogram Gradient Boosting Regression (HGBR) approach provides significant advantages for regression tasks, demonstrating robust predictive capabilities due to its adaptive handling of diverse data types and structures (Liu, 2024). Likewise, Naeem et al. discuss HGB's capacity for producing accurate predictions, resulting from its unique ability to iteratively refine models based on histogram approximations of data distributions (Setyarini et al., 2024). The combination of speed and performance makes HGB particularly suitable for educational contexts, where large datasets containing both continuous and categorical variables are common.

Moreover, HGB's ability to scale effectively with large datasets further solidifies its role as a strong candidate for deployment in educational settings. Liu's research validates this approach by showcasing its superiority over traditional models, highlighting that HGB outperforms linear regression techniques or other classification techniques in predicting student performance, thus enabling timely intervention strategies for at-risk students (Liu, 2024). The framework's efficient resource utilization is particularly crucial in educational institutions where computational resources may be limited and timely results are required for decision-making.

Hybrid models, combining strengths of multiple algorithms (e.g., MLP with decision trees), often achieve the highest accuracy (Siddique et al., 2021; Ayienda et al., 2021). However, they come with increased complexity and reduced interpretability, making them challenging to implement and explain in non-technical educational environments.

Ultimately, while no single model is universally superior, histogram-based gradient boosting and MLP emerged as the final selected models in this study due to their combined strengths in predictive accuracy, data adaptability, and computational efficiency. A detailed comparison of these techniques is presented in the table below.

Table 1: Comparison of Machine Learning Techniques for Student Performance Prediction

Technique	Strengths	Weaknesses
Multilayer Perceptron (MLP)	High accuracy, handles complex data relationships	Low interpretability, requires large datasets
Random Forest (RF)	Reduces overfitting, interpretable feature importance	Less effective for imbalanced data
XGBoost / Gradient Boosting	Strong predictive power, reduces bias and variance	Computationally expensive, complex to tune
Histogram-Based Gradient Boosting	Efficient for large datasets, handles missing values, supports categorical features	Less interpretable, requires parameter tuning
Decision Trees (DT)	Easy to interpret, fast training	Prone to overfitting, low generalization
Support Vector Machines (SVM)	Effective in high-dimensional space	Poor scalability, limited interpretability
K-Nearest Neighbors (KNN)	Simple implementation, no training phase	Sensitive to noise, inefficient on large datasets
Naïve Bayes (NB)	Scalable, handles large categorical datasets	Assumes feature independence, may reduce accuracy
Hybrid Models	Combine benefits of multiple models, high accuracy	Increased complexity, difficult to interpret

2.5 Algorithmic Bias and Fairness in AI-Powered Prediction Models

Algorithmic bias refers to systematic and repeatable errors in AI predictions that unfairly favour or disadvantage certain groups of individuals, often reflecting existing social inequalities embedded within training data. This bias may originate from the data itself, which may exhibit historical inequities, or from the algorithms used, potentially leading to biased outcomes during prediction (Baker and Hawn, 2021). In educational contexts, algorithmic bias can result in unequal resource allocation, misclassification, and support interventions that perpetuate disparities in academic achievement, particularly among marginalised student groups.

As Baker and Hawn (2021) highlight, the manifestation of algorithmic bias in educational algorithms can adversely affect underrepresented populations, contributing to ongoing achievement gaps rather than mitigating them. These issues are especially concerning in high-stakes settings such as education, where predictive models may influence critical decisions related to academic progression and access to resources.

2.5.1 Sources of Bias in Educational Data

Bias in educational AI systems can arise from multiple sources. Imbalanced datasets that underrepresent certain demographic groups can skew predictions (Fazil et al., 2024). Historical academic records often encode past inequities, while proxy variables such as geographic location or attendance rates may indirectly reflect sensitive characteristics like race or socio-economic status. As noted by Fazil et al. (2024), predictive models trained on historically biased data may misrepresent the abilities of disadvantaged students, reinforcing existing structural disparities. Similarly, Yagci (2022) found that machine learning models built on biased data tend to reinforce educational inequalities, disproportionately impacting underprivileged groups.

2.5.2 Defining and Evaluating Fairness in AI Models

Fairness in AI is commonly defined as the equitable treatment of individuals or groups, irrespective of protected attributes such as race, gender, or socio-economic status (Mitchell et al., 2021). In the context of educational predictive modelling, several fairness criteria are employed to evaluate model equity. These include concepts such as demographic parity, individual fairness, and group fairness, which collectively help assess whether predictions are fairly distributed across diverse student populations status (Mitchell et al., 2021).

To operationalise these principles, researchers apply a range of fairness metrics such as the disparate impact ratio, statistical parity difference, and equal opportunity difference to detect and quantify disparities in model outcomes across demographic groups. The disparate impact ratio measures differences in favourable outcomes between protected and unprotected groups, providing a signal for potential discrimination (Patrikar et al., 2023). Statistical parity difference captures variations in the probability of receiving positive outcomes, while equal opportunity difference ensures consistent true positive rates across groups, fostering equitable access to beneficial results (Mehrabi et al., 2021).

However, as noted by Mehrabi et al. (2021) and Bhanot et al. (2021), the implementation of fairness assessments often hinges on access to sensitive demographic attributes. While these data are essential for identifying and mitigating biases, their use introduces ethical and privacy challenges particularly in domains such as education, healthcare, and criminal justice (Akgün et al., 2023; Chauhan et al., 2023). The absence of such data can obscure systemic disparities, yet including them requires stringent data protection measures to safeguard individual privacy and prevent misuse (Bhanot et al., 2021).

Despite these complexities, ensuring fairness in predictive modelling remains a critical priority especially in high-stakes environments where algorithmic decisions can reinforce structural inequalities (Hickey et al., 2020; Pereira et al., 2021). This calls for the development of robust, transparent, and ethically sound evaluation frameworks. Emerging techniques such as differential privacy, fair representation learning, and the use of synthetic or anonymised datasets present viable strategies to balance fairness and privacy (Pereira et al., 2024; Barbierato et al., 2022). Ultimately, fairness in algorithmic systems extends beyond technical solutions and demands a sustained, interdisciplinary commitment to ethical accountability and social justice (Weerd, 2024).

2.5.3 Mitigating Algorithmic Bias

The complexity of defining fairness in educational contexts has prompted researchers to develop a range of fairness-aware machine learning techniques. Gupta and Khosla (2021) explore methods such as adversarial debiasing, reweighting of training samples, and fairness-aware loss functions, each designed to reduce disparities during training. Kamiran and Calders (2012) propose a complementary set of approaches: pre-processing methods to cleanse biased datasets, in-processing techniques that embed fairness constraints during training, and post-processing methods that adjust model predictions to minimise discriminatory outcomes. While effective, these methods often involve trade-offs with model performance.

To mitigate bias, strategies are implemented at various stages of the machine learning pipeline. Pre-processing techniques such as Synthetic Minority Over-sampling Technique (SMOTE) and re-weighting help address class imbalance and data skewness (Liu, 2024). In-processing strategies, including fairness-aware algorithms and adversarial debiasing, incorporate equity constraints directly into model training (Mitchell et al., 2021). Post-processing approaches modify the model's predictions to ensure fairer outcomes across groups. These combined strategies improve fairness without severely compromising model accuracy.

2.5.4 Ethical and Practical Considerations

From an ethical perspective, developers and stakeholders have a responsibility to uphold fairness, transparency, and accountability. XAI techniques can enhance transparency by allowing stakeholders to understand how specific features influence predictions (Gupta et al., 2024). However, the trade-off between fairness and accuracy remains a critical issue. Fairness-aware models often yield more equitable outcomes but may sacrifice some degree of predictive performance. This makes human oversight and domain-specific knowledge essential in balancing these competing objectives (Mienye and Sun, 2022).

2.5.5 Relevance to This Study

This study actively addresses algorithmic bias through a comprehensive framework. Sensitive features such as race, gender, and socio-economic status were either excluded or transformed to reduce potential bias. Fairness metrics, including statistical parity difference and equal opportunity difference, have been used to evaluate model performance across groups. Additionally, XAI techniques have been applied to ensure interpretability and transparency in feature contributions. These measures reflect a commitment to ethical and fairness-aware AI in education.

2.6 Data Privacy and Security Concerns

The use of AI in education relies on vast amounts of student data, including academic records, behavioural metrics, and even socio-economic factors. This raises concerns about data privacy and security, particularly regarding how student information is collected, stored, and shared.

The General Data Protection Regulation (GDPR) and similar legislation set strict guidelines for handling personal data, requiring educational institutions to ensure informed consent, data minimization, and secure data storage. However, compliance with these regulations remains challenging, particularly as AI models require extensive datasets for training. Li and Liu (2020) highlighted the risks of data breaches and unauthorized access to student records, which could lead to identity theft, profiling, or misuse of sensitive information.

Privacy-preserving machine learning (PPML) techniques have been proposed to address these challenges. Federated learning, a technique that allows models to be trained across multiple decentralized devices without transferring raw data, has gained attention for its ability to protect student privacy. Abadi et al. (2016) introduced differential privacy as another solution, where noise is added to data during the training process to prevent the identification of individual students while maintaining overall model utility.

Despite these technological advances, data privacy remains a pressing ethical issue in AI-driven education. Educational institutions must establish clear policies on data governance, ensure student consent is obtained transparently, and invest in secure data infrastructure to prevent misuse.

2.6.1 Ethical AI Governance and Policy Recommendations

To ensure the responsible use of AI in education, there is a growing need for ethical AI governance frameworks. Various policy organizations, including UNESCO and the European Union, have proposed guidelines for ethical AI implementation.

The UK government's Centre for Data Ethics and Innovation (CDEI) has also emphasized the importance of ethical AI adoption in education. Their 2021 report recommended that AI systems used in schools undergo regular audits to assess fairness, transparency, and accountability. Additionally, they highlighted the need for teacher training in AI literacy to ensure that educators understand how AI tools function and can critically assess their outputs.

Furthermore, researchers such as Cowls and Floridi (2018) advocate for value-sensitive design (VSD) in AI systems, ensuring that ethical considerations are embedded at the design stage rather than addressed as an afterthought. By involving educators, students, and policymakers in the AI development process, institutions can create technology that aligns with educational values and priorities.

The ethical deployment of AI in education requires careful consideration of bias mitigation, explainability, data privacy, and human oversight. While machine learning has the potential to revolutionize student performance prediction and personalized learning, its implementation must be guided by principles of fairness, transparency, and accountability. This research focuses on integrating fairness-aware ML techniques, developing interpretable models, and establishing clear policies on data governance. By prioritizing ethical considerations, educational institutions can ensure that AI serves as a force for equity and innovation rather than reinforcing existing disparities.

2.6.2 The Role of Human Oversight in AI Systems

AI should not replace human decision-making in education; instead, it should function as a supportive tool that complements the expertise of educators and policymakers (Luckin et al., 2020). This perspective underscores the necessity of maintaining human oversight at the heart of all AI-assisted decision-making processes. A study by Luckin et al. (2020) emphasized the importance of integrating AI with teacher judgment rather than delegating final decisions on student performance and progression to algorithms. They argue that AI should augment human insight by offering recommendations, not deterministic classifications while ensuring educators retain the autonomy to consider contextual factors and override automated outputs when appropriate.

Similarly, Agarwal et al. (2024) stress that the ethical deployment of AI in education requires positioning technology as an adjunct to, rather than a replacement for, human judgment. Their concept of the "educational aptness principle" reinforces the need to align AI implementation with ethical standards tailored to educational environments. The European Commission’s Ethics Guidelines for Trustworthy AI (2019) also echo this human-centric approach. These guidelines advocate for transparency, accountability, and the right to contest AI-generated decisions, recommending that educational institutions establish robust frameworks to evaluate the effects of AI on student outcomes. Such safeguards help ensure that technological tools enhance rather than diminish educational opportunities.

The need for ethical awareness in AI education is further supported by Sanusi and Olaleye (2022), who argue that future technologists must be equipped with a strong ethical foundation to fully understand the societal implications of their innovations. Viewing AI as a purely technical tool risks overlooking the broader ethical landscape, especially in education, where learner development is deeply personal and context-sensitive. Holmes et al. (2021) similarly emphasize the importance of continuously evolving ethical frameworks to keep pace with rapid developments in educational technologies.

Jobin et al. (2019) contribute to this discourse by highlighting that building public trust in AI requires sustained critical scrutiny particularly in education, where algorithmic decisions can profoundly shape learners' futures. Although AI has the potential to enhance educational methodologies, Leddy and Creanor (2024) caution that it must not displace the irreplaceable human elements of teaching, such as empathy, adaptability, and contextual understanding.

In conclusion, as AI systems become more embedded in educational settings, it is essential to ensure they support rather than supplant human decision-making. This balance requires ongoing dialogue around ethical practice extending beyond technical efficiency to encompass social responsibility, accountability, and human dignity. By foregrounding the role of educators and maintaining human agency, stakeholders can ensure that AI technologies contribute meaningfully to equitable and effective education for all.

2.7 Evaluation and Validation Practices in Educational AI Models

Evaluation and validation techniques are crucial for assessing the performance, fairness, and generalisability of AI models in educational contexts. These methods form the backbone of responsible AI development, ensuring that models perform effectively on historical data while aligning with educational goals and upholding ethical standards. Traditionally, model evaluation in educational settings has focused on accuracy-based metrics such as accuracy, precision, recall, and F1-score. While these metrics are essential for measuring technical performance, they are often insufficient on their own particularly in applications where decisions have direct and lasting impacts on students' academic futures (Roshanaei et al., 2023).

Accuracy, defined as the proportion of correct predictions out of all predictions made, is the most commonly used metric. In tasks such as predicting student outcomes or identifying at-risk learners, it provides a straightforward assessment of model correctness. However, accuracy can be misleading in cases where datasets are imbalanced. For example, in a dataset where 90 percent of students are not at risk of dropping out, a model that always predicts "no risk" would achieve 90 percent accuracy while completely failing to identify the students who actually need support (Currie, 2019).

To address this limitation, precision and recall provide more nuanced insights. Precision measures the proportion of true positive predictions among all positive predictions made by the model. In education, this could reflect how many of the students flagged as "at risk" truly require intervention. This metric becomes particularly important when the cost of false positives is high, such as when a student is incorrectly identified as failing, potentially leading to unnecessary interventions or anxiety (Li et al., 2023). Recall, in contrast, measures how many of the actual positive cases the model correctly identifies. It is critical in scenarios in which failing to detect at-risk students could result in serious academic consequences (Chen et al., 2020). Considering that precision and recall often exist in tension, the F1-score, which is the harmonic mean of the two, is especially useful. It provides a balanced view of the model's performance, particularly in cases where both false positives and false negatives carry significant weight (Roshanaei et al., 2023).

While these accuracy-based metrics are foundational, several scholars caution against their exclusive use. Zeng et al. (2016) and Jobin et al. (2019) argue that reliance on traditional metrics alone can obscure important issues such as biased data distributions or unfair decision-making processes that disproportionately impact marginalised groups. Şahin et al. (2023) similarly warn that models trained on biased historical data can perpetuate inequality, particularly when socio-demographic attributes are not ethically managed. Blow et al. (2024) add that such biases can remain undetected without intentional fairness assessments during model development.

In response to these challenges, contemporary research advocates for multi-dimensional evaluation frameworks. Rattanaphan and Briassouli (2024) propose combining standard accuracy metrics with fairness indicators, including equal opportunity difference, demographic parity, and disparate impact ratio. These indicators help to evaluate whether AI-generated outcomes are distributed equitably across different student demographics (Theodorou et al., 2022). Even highly accurate models may underperform for specific subgroups, such as students from economically disadvantaged backgrounds, thereby producing unjust outcomes despite seemingly strong performance (Jobin et al., 2019). Evidence from Dressel and Farid (2018) supports this concern, demonstrating that models with high technical accuracy may still produce biased predictions when evaluated across diverse populations.

Fairness-aware evaluation frameworks are increasingly adopted in the educational AI field. Bellamy et al. (2019) and Mitchell et al. (2021) both recommend integrating ethical and statistical assessments into model evaluation. The model cards framework developed by Mitchell et al. provides a structured method for documenting a model's performance, fairness, and known limitations, promoting transparency and accountability (Theodorou et al., 2022).

In addition to fairness metrics, validation techniques are essential for assessing model generalisability. K-fold cross-validation, which partitions the dataset into multiple training and testing sets, helps ensure robust performance estimates. Stratified k-fold cross-validation improves this process by preserving class distributions across folds, making it particularly useful for educational datasets that are often imbalanced (Seo et al., 2021). Nested cross-validation further reduces the risk of overfitting during hyperparameter tuning by separating model selection from final performance evaluation (Pucchio et al., 2022). Hold-out validation using temporal or demographic splits is also important for assessing a model's ability to generalise across different academic years or student groups.

Innovative approaches such as the train-then-mask method introduced by Ghili et al. (2019) help reveal model dependence on protected attributes. This technique involves training the model with all features and then masking sensitive attributes at prediction time, allowing researchers to assess whether predictions are unduly influenced by those attributes (Roshanaei et al., 2023). Other bias mitigation strategies include adversarial debiasing and the use of fairness constraints during model training, which proactively reduce bias by embedding fairness objectives into the learning process (Yaseliani et al., 2024).

Equally important is human-in-the-loop validation, which incorporates the judgement of educators, domain experts, and institutional decision-makers. This collaborative approach ensures that AI outputs are contextually appropriate and aligned with educational values and goals, particularly in edge cases where automated reasoning may be insufficient (Chen et al., 2020).

Below is a table summarising the identified evaluation and validation techniques identified from the literature.

Table 2: Accuracy-Based Evaluation Metrics and Validation Techniques in Educational AI

Category	Technique	Purpose in Educational AI	Mathematical Formula	Citation
Accuracy Metrics	Accuracy	Measures overall correctness of predictions	Accuracy = (TP + TN) / (TP + FP + TN + FN)	Roshanaei et al., 2023
	Precision	Evaluates correctness of positive predictions (e.g., students flagged as at-risk)	Precision = TP / (TP + FP)	Li et al., 2023
	Recall	Assesses model's ability to detect all actual positives (e.g., identifying all at-risk students)	Recall = TP / (TP + FN)	Chen et al., 2020
	F1-Score	Balances precision and recall, ideal for imbalanced classes	F1-Score = 2 * (Precision * Recall) / (Precision + Recall)	Roshanaei et al., 2023
Validation Techniques	K-Fold Cross-Validation	Estimates general performance across data splits	No specific formula; involves partitioning the data into k subsets	Seo et al., 2021
	Stratified K-Fold Cross-Validation	Preserves class distribution across folds; useful in imbalanced educational datasets	Similar to k-fold but preserves class distribution in each fold	Seo et al., 2021
	Nested Cross-Validation	Prevents overfitting during hyperparameter tuning	Nested CV = Outer CV for performance, Inner CV for hyperparameter tuning	Pucchio et al., 2022

	Hold-Out Validation	Evaluates model on a separate temporal or demographic dataset	Split data into training and testing sets once	Currie, 2019
	Bootstrapping	Estimates confidence intervals for performance metrics	Resample with replacement and compute metrics repeatedly	Roshanaei et al., 2023
Fairness Techniques	Equal Opportunity Difference	Evaluates whether true positive rates are equal across groups	TPR_difference = TPR_groupA - TPR_groupB, where TPR represents True Positive Rates	Theodorou et al., 2022
	Demographic Parity	Checks if positive outcomes are equally distributed among groups	P(Y=1|A=0) = P(Y=1|A=1)	Theodorou et al., 2022
	Disparate Impact Ratio	Compares outcome rates for protected vs. unprotected groups	Disparate Impact = P(Y=1|A=1) / P(Y=1|A=0)	Theodorou et al., 2022
Bias Testing	Train-Then-Mask (Ghili et al.)	Tests model reliance on protected attributes	Compare model predictions with and without protected features	Ghili et al., 2019
	Adversarial Debiasing	Actively reduces bias during model training	Involves training a model to remove bias during learning using adversarial loss	Yaseliani et al., 2024
	Fairness Constraints	Integrates fairness into loss functions during training	Applies constraints such as fairness regularization in the loss function	Yaseliani et al., 2024
Validation Techniques	Leave-One-Out Cross-	Each instance is used once as	Each fold = 1 instance as test,	Seo et al., 2021

	Validation (LOOCV)	a test set; useful for very small datasets	remaining n-1 as training; repeat n times	
Validation Techniques	Time Series Split	Used when data is sequential (e.g., student performance over terms); maintains temporal order	Sequential splits; no formula but involves forward chaining for training and test sets	Currie, 2019

In conclusion, while accuracy, precision, recall, and F1-score are indispensable tools for evaluating AI performance, they must be complemented by fairness metrics, advanced validation methods, and human oversight. Only through a comprehensive evaluation approach that integrates these elements can educational AI systems be made both effective and equitable, ensuring that they benefit all learners without reinforcing existing disparities (Roshanaei et al., 2023).

2.8 Exploring Explainable AI techniques for predictive modelling in Education

Explainable AI pertains to a collection of procedures and techniques that strive to offer a lucid and comprehensible explanation for the decisions produced by AI and machine learning models. By incorporating an explainability layer onto these models, Data Scientists and Machine Learning practitioners can construct more reliable and clear systems to support many stakeholders, including developers, regulators, and end-users (Zednik & Boelson, 2022).

Accurate data is essential in the field of machine learning prediction. With predictions produced by AI models, we often depend on intricate computer models to provide us with results, but we are uncertain about the precise methodology employed by these models to generate their outputs. In a study conducted by Zhao (2021), it was found that explainable model tools or procedures play a crucial role in enhancing the understanding of their functioning. The absence of openness in the acquired data gives rise to apprehensions regarding the capacity to rely on and comprehend all the models, as well as the biases, fairness, and ethical ramifications, specifically in high-stakes applications.

In response to these difficulties, the discipline of XAI has arisen, with the specific aim of elucidating the internal mechanisms of black box models.

2.8.1 Defining Explainability, Interpretability and Transparency

In the realm of predictive modelling, particularly within artificial intelligence, the concepts of explainability, interpretability, and transparency are essential yet distinct dimensions that collectively determine the usability, accountability, and trustworthiness of AI systems. These attributes not only support model comprehension for developers and end-users but also underpin ethical deployment, particularly in sensitive domains such as education.

Explainability refers to the degree to which an AI model's internal mechanics and decision-making processes can be articulated and understood by human stakeholders (Patrikar et al., 2023). It is often associated with post-hoc techniques that provide insight into otherwise opaque models, such as deep neural networks or ensemble models. Explainability addresses the question: why did the model make a specific prediction? It involves generating human-readable justifications that help various stakeholders such as teachers, policy-makers, or students understand outcomes and trust the AI system (Patrikar et al., 2023). According to Merry et al. (2021), explainability must be tailored to the contextual needs of the audience, as different stakeholders may require different levels or types of information. This audience-sensitive approach highlights the situational nature of explanation: what may be sufficiently explainable for a data scientist may be entirely opaque to a school administrator or parent. Therefore, effective explainability requires methods that are not only technically accurate but also contextually relevant and accessible.

Interpretability, while often used interchangeably with explainability, is more narrowly focused on the extent to which a human can directly understand the internal logic and structure of the model without needing additional interpretive tools (Mehrabi et al., 2021). It answers the question: can the user intuitively grasp how the model works? Interpretability is typically associated with simpler, inherently transparent models such as linear regression, decision trees, or rule-based classifiers sometimes referred to as glass box or white box models. These models allow users to trace input-output relationships and understand how specific features contribute to predictions. As noted by Mehrabi et al. (2021), interpretable models are particularly valuable in high-stakes settings, as they provide clarity and reassurance to stakeholders, reduce reliance on post-hoc rationalizations, and are less likely to mask biases. In contrast, complex models like random forests or deep neural networks despite their high predictive power are often labelled as black boxes due to their low interpretability. While such models may offer improved performance, their opaque structure makes it difficult to diagnose errors, identify sources of bias, or justify decisions.

Transparency, distinct yet complementary, refers to the openness and visibility of the entire AI development pipeline, including data collection practices, feature selection, model architecture, training procedures, and deployment protocols (Bhanot et al., 2021). It is not limited to the model's mechanics but encompasses the full lifecycle of AI system development. Transparency is about ensuring that stakeholders can access and scrutinize how data is collected, what assumptions are embedded in the model, and how predictions are generated. Bhanot et al. (2021) highlight that transparency is increasingly critical in the face of growing model complexity, calling for documentation standards and tools that allow developers, auditors, and users to trace decisions back to specific model components or training data sources. Transparent systems support reproducibility, regulatory compliance, and informed decision-making, all of which are essential for building public trust in AI.

Together, explainability, interpretability, and transparency form a triad that supports not only technical robustness but also ethical legitimacy. They are particularly vital in domains where AI predictions have significant social or personal consequences. Akgün et al. (2023) emphasize that without these qualities, AI models risk being deployed in ways that are opaque, unaccountable, and potentially discriminatory. In high-stakes contexts, users must be able to interrogate AI outcomes, identify sources of error or bias, and crucially contest decisions when necessary. This is especially true when models influence decisions related to student performance, eligibility for services, medical diagnoses, or legal judgments.

Recent frameworks for explainable AI, such as those discussed by Suthaharan Chauhan et al. (2023), aim to operationalize these principles by incorporating both technical and human-centred strategies. These frameworks often combine algorithmic techniques like SHAP, LIME, and counterfactual reasoning with ethical design principles, user-centred interfaces, and participatory design methodologies. By enabling users to explore the rationale behind model predictions and by facilitating dialogue between technical and non-technical stakeholders, these approaches help bridge the gap between AI sophistication and societal expectations.

In summary, explainability, interpretability, and transparency are critical to building trustworthy AI systems. Each concept addresses a different layer of understanding from the mechanics of the model to the visibility of its development and the accessibility of its outputs. Together, they support responsible AI practices by ensuring that models are not only effective but also fair, understandable, and aligned with the values of the communities they serve. Effective communication and comprehension of model predictions hinge on these principles being rigorously implemented, empowering users to engage with AI technologies confidently and responsibly.

2.8.2 XAI Techniques for Interpreting Complex Models

This section examines the interpretability of complex machine learning predictions using prominent XAI techniques and their adaptations for use in the context of education. These include LIME, PDP, and ALE. Each technique brings unique advantages in illuminating how input features influence predictive outcomes, laying a foundation for responsible, transparent, and equitable educational AI systems.

2.8.2.1 SHAP (Shapley Additive Explanations)

SHAP, rooted in cooperative game theory, computes Shapley values to quantify the marginal contribution of each input feature to a given prediction. It offers both global and local interpretability: at the global level, it identifies which features are most influential across all predictions; at the local level, it explains how individual features impact a specific prediction. In educational contexts, SHAP is particularly valuable for fairness assessments and bias detection. For example, it can reveal if a model systematically overemphasizes socio-economic status in predicting academic outcomes. Such insights allow educators and developers to identify unintended biases and make corrective adjustments. SHAP’s visual tools such as force plots, dependence plots, and summary plots enable complex explanations to be communicated effectively to non-technical stakeholders (Patrikar et al., 2023; Mehrabi et al., 2021).

2.8.2.2 LIME (Local Interpretable Model-Agnostic Explanations)

LIME approximates complex models using interpretable, locally faithful surrogate models, often linear regressions or decision trees. By perturbing the input data around a specific instance and analyzing the corresponding output changes, LIME generates simplifies explanations for individual predictions. Its model-agnostic design ensures flexibility, making it applicable across a wide range of predictive models. In educational settings, LIME is particularly useful for interpreting decisions about individual students such as risk of dropout or projected grade performance where transparency at the individual level is crucial for justifying interventions. While LIME lacks global model interpretability, its strength lies in supporting personalized, case-specific decision-making (Bhanot et al., 2021; Akgün et al., 2023).

2.8.2.3 Partial Dependence Plots (PDP)

PDPs are visual tools that illustrate the marginal effect of a single feature on the predicted outcome by averaging the influence of all other features. This technique helps uncover how continuous variables such as prior attainment scores or parental involvement correlate with predicted educational performance. While PDPs are useful for communicating trends to non-technical users, their primary limitation is the assumption of feature independence, which may not hold in many educational datasets where variables are often correlated. Nonetheless, PDPs can inform strategic decisions regarding resource allocation or curriculum design (Chauhan et al., 2023; Hickey et al., 2020).

2.8.2.4 Accumulated Local Effects (ALE)

ALE plots address key limitations of PDPs by accounting for feature interactions and correlations. ALE estimates the average effect of a feature within its value intervals, based on the local structure of the data distribution. This makes it more reliable and unbiased when features are correlated, an important consideration in education, where variables such as attendance, engagement, and socio-economic background are often interrelated. ALE provides a refined understanding of how such features jointly influence predictions and is well-suited for applications involving complex student performance data (Pereira et al., 2021; Pereira et al., 2024).

2.8.2.5 Integrating XAI Techniques

Combining SHAP, LIME, PDP, and ALE provides a comprehensive and robust interpretability framework. Each technique contributes a unique perspective ranging from individual to global explanations and from linear effects to interaction-aware analyses. Integrating these methods facilitates a multidimensional understanding of model behavior, helping educators, administrators, and policymakers interpret AI outputs with greater confidence. This holistic approach supports responsible AI deployment by fostering transparency, enabling model validation, and guiding ethical decision-making in education (Barbierato et al., 2022).

2.8.3 Related Work and Application of XAI Techniques in Education

Although XAI methods like SHAP and LIME have gained traction in domains such as healthcare, finance, and law, their adoption in education remains emergent. Much of the early application of SHAP has focused on interpreting black-box models such as neural networks and support vector machines. However, recent contributions by Amarasinghe et al. have extended SHAP for use in educational prediction tasks, demonstrating its effectiveness in making deep learning models interpretable in domains such as student performance prediction, behavior modeling, and early intervention planning (Weerd, 2024).

In practice, SHAP has been employed to interpret models that predict student dropout risks, learning progression, or intervention outcomes. By identifying and ranking the influence of features such as attendance, prior achievement, and socio-economic status, SHAP enables fairness-aware diagnostics and promotes informed, data-driven educational decisions. Its dual utility in technical analysis and ethical evaluation makes it a critical tool for AI governance in education (Agarwal et al., 2024).

While LIME has not yet been widely adopted in education research, its strengths in real-time interpretability offer promise for use in intelligent tutoring systems, adaptive learning environments, and personalized feedback platforms. By providing immediate, instance-specific explanations, LIME allows educators and students to understand and respond to AI-generated insights (Sanusi and Olaleye, 2022).

Similarly, PDP and ALE have shown utility in exploratory analysis, helping researchers visualize how key predictors affect outcomes. However, broader implementation of these techniques is needed to standardize their role in model interpretation and policy auditing within educational AI systems (Holmes et al., 2021; Jobin et al., 2019).

The integration of XAI techniques into educational machine learning systems is not merely a technical enhancement but a moral imperative. SHAP, LIME, PDP, and ALE serve as essential tools for interpreting complex models, enabling transparency, fairness, and accountability. These techniques can facilitate stakeholder trust and inform action, ensuring that AI supports pedagogical goals while safeguarding against bias and opacity. As educational systems increasingly incorporate predictive technologies, we believe explainability will remain central to aligning innovation with equity and ethical responsibility.

2.8.4 Evaluating XAI techniques

While XAI techniques offer promising strategies for enhancing the transparency and accountability of AI systems, several challenges hinder their effective implementation, especially in high-stakes domains like education. A primary concern is scalability. Techniques such as SHAP, although powerful, often become computationally intensive and impractical when applied to large datasets or complex architectures like deep neural networks (Patrikar et al., 2023). Equally pressing is the lack of standardized metrics to evaluate the quality of explanations. As noted by Mehrabi et al. (2021), this absence of consensus leads to inconsistent evaluations across studies, making it difficult to compare results or establish best practices for selecting and applying XAI techniques.

Evaluation methodologies currently span a wide spectrum, ranging from quantitative, automated techniques to user-centric and counterfactual-based approaches. However, many of these methods fall short in meeting the specific needs of educational contexts, where clarity, accessibility, and contextual relevance are paramount.

For instance, Radingoana (2023) explored the evaluation of textual explanations generated by XAI algorithms using automated quantitative metrics such as BLEU (Bilingual Evaluation Understudy Score), METEOR (Metric for Evaluation of Translation with Explicit Ordering), and CIDEr (Consensus-based Image Description Evaluation). While these metrics are well-established in the domain of Natural Language Processing (NLP) for evaluating machine-generated text against human references, their applicability in educational XAI is limited. They tend to prioritize sentence-level similarity over the semantic clarity or contextual appropriateness necessary for educational stakeholders such as teachers and administrators (Bhanot et al., 2021; Akgün et al., 2023).

Loef (2022) applied SHAP and LIME to detect credit card fraud within deep learning and random forest models. However, the study emphasized model performance metrics such as accuracy, recall, sufficiency, and F1-score, which, although important for evaluating predictive performance, are less suitable for assessing the quality and usefulness of explanations generated by XAI tools (Chauhan et al., 2023). Similarly, Kakogeorgiou and Karantzalos (2021) explored the application of XAI techniques in remote sensing and multi-label classification tasks. They introduced a novel trust metric designed to evaluate interpretability techniques independently of specific machine learning algorithms or tasks. While their goal was to promote intuitive and accurate decision-making, the approach faced challenges related to computational complexity and scalability in practical applications (Hickey et al., 2020).

Alternative approaches have focused on subjective, user-centric assessments. Kotecha (2021), for example, employed questionnaires to evaluate users' perceptions of model interpretability. While this method captures valuable qualitative insights, it is inherently dependent on subjective judgments, which can vary significantly among individuals and limit reproducibility (Pereira et al., 2021). In response to these limitations, Torres (2022) called for more impartial and inclusive evaluation frameworks for XAI techniques, particularly SHAP and LIME. The study emphasized that the technical complexity of many XAI tools can alienate non-expert users, thus reducing trust and hindering adoption in domains like education (Pereira et al., 2024).

Puram (2023) proposed a framework focused on evaluating interpretability through user comprehension, aiming to align explanations with human understanding. While this contribution highlights the importance of user-centred design in XAI, it lacks a systematic methodology for evaluating and comparing specific techniques such as SHAP, LIME, PDP, and ALE. The absence of robust, quantitative evaluation tools remains a significant gap in the field (Barbierato et al., 2022).

More recently, Liu (2024) introduced an innovative methodology for assessing the faithfulness of XAI explanations through counterfactual reasoning. This approach incorporates two primary metrics: validity and proximity. Validity evaluates whether the features highlighted in an explanation meaningfully influence the model's output, while proximity assesses the minimal changes needed in input features to alter the model's predictions. These measures are synthesized into a comprehensive Counterfactual Evaluation Score (CES), intended to gauge how well explanations reflect the model's actual decision-making processes (Weerd, 2024).

While Liu's CES framework offers a promising direction for systematic XAI evaluation, it is not without limitations. The computational burden associated with generating counterfactual examples and computing CES is substantial, particularly when applied to high-dimensional or large-scale datasets (Agarwal et al., 2024). Moreover, CES may not generalize across different model types, especially those involving complex non-linear interactions. Another limitation arises from the assumption of feature independence during counterfactual generation, which may oversimplify the interdependencies inherent in real-world educational data (Sanusi and Olaleye, 2022). Additionally, while CES is designed to evaluate faithfulness, it does not directly address interpretability from the user's perspective, which may hinder its applicability among educators or policymakers lacking technical expertise (Holmes et al., 2021).

In conclusion, while recent advancements such as CES present more structured frameworks for evaluating XAI, their complexity and underlying assumptions limit practical deployment in educational settings. The lack of standardized, domain-sensitive evaluation approaches continues to be a barrier to the broader adoption of XAI. Given the importance of interpretability, especially in education where stakeholders must make informed decisions based on AI outputs, future research must focus on developing inclusive, scalable, and pedagogically grounded evaluation methodologies. These should combine objective performance with human-cantered criteria to ensure that XAI techniques are not only technically valid but also understandable, trustworthy, and actionable across diverse educational environments (Jobin et al., 2019).

2.9 Predictive Modelling of GCSE Outcomes in the UK Context

While extensive research has examined student performance prediction using machine learning across various international contexts, relatively few studies have focused specifically on the United Kingdom's General Certificate of Secondary Education (GCSE) system. Nevertheless, a growing body of UK-based research has begun to bridge this gap by applying AI-driven methods to support educational forecasting, intervention strategies, and policymaking in secondary education.

A seminal study by Anders et al. (2020) employed machine learning models to predict GCSE outcomes using large-scale administrative data from the National Pupil Database (NPD). Their models incorporated both academic and non-academic features such as socioeconomic background and school-level characteristics to enhance predictive performance. Importantly, their findings demonstrated that including these non-academic variables significantly improved accuracy but also introduced risks of perpetuating bias. Without appropriate fairness constraints, predictive models could reinforce existing socioeconomic and ethnic disparities in educational outcomes (Arowosegbe et al., 2024).

In a related effort, Holloway and Gulliver (2021) examined how school-level predictors, including funding per pupil, teacher-to-student ratios, and regional deprivation indices, influenced GCSE performance. Their work underscored the critical role of institutional and environmental factors in shaping academic outcomes and emphasised the need for educational data models that account for both individual and contextual variables.

More recently, Denes (2023) applied ensemble machine learning techniques namely Random Forests and Gradient Boosting to predict subject-level GCSE grades in a selective independent school in England. The study utilised features such as prior attainment, attendance, and demographics. Despite its limited scope and focus on a high-achieving student population, the research achieved over 80 percent accuracy in some subjects. This reinforces the potential of machine learning to contribute meaningfully to school-level analytics and decision-making.

Collectively, these studies demonstrate the promise of AI in enhancing educational planning and identifying at-risk learners. However, several persistent challenges continue to constrain the broader application of these models in the UK context. These include:

- Restricted access to sensitive or personal data (such as family income and behavioural history) due to ethical and legal considerations.
- Variability in curricula and grading standards across schools, which complicates model generalisability.
- Limited transparency and interpretability of many high-performing machine learning models, which hinders their adoption by educators and decision-makers.

In response to these challenges, recent research has begun to advocate for the integration of explainable artificial intelligence (XAI) methods into predictive modelling pipelines. Human-centric explainability frameworks, such as those proposed by Maity and Deroy (2024), argue that transparent models are essential for trust, accountability, and practical deployment. This view is supported by Rainey et al. (2021), who emphasise that explainability enhances the legitimacy and usefulness of AI tools in high-stakes environments like education. Complementing these perspectives, Arrieta et al. (2020) and Elhage et al. (2021) provide detailed taxonomies of explainability challenges and call for greater integration of XAI into educational data systems.

Building on this foundational work, the present study aims to combine fairness-aware machine learning with interpretable modelling techniques to predict GCSE performance. By embedding transparency and stakeholder-specific explanations into the modelling process, this research seeks to improve both the usability and ethical robustness of AI applications in secondary education. In doing so, it contributes to the growing call for data-driven educational strategies that are accurate, actionable, and aligned with principles of equity and responsible innovation.

2.10 Chapter Summary and Synthesis

This chapter has presented a comprehensive review of the literature on student performance prediction, highlighting the integration of ML and explainable artificial intelligence XAI within educational contexts. The review critically examined the multidimensional factors influencing academic outcomes, including demographic, socio-economic, behavioural, psychological, and technological variables. These diverse influences underscore the importance of adopting holistic prediction models rather than relying on isolated factors such as gender or socio-economic status.

While ML models, particularly ensemble techniques and deep learning approaches like MLPs, have demonstrated high predictive accuracy, their limited interpretability poses a significant barrier to adoption in education. These models often operate as "black boxes," offering little insight into how predictions are generated. This lack of transparency raises concerns around trust, accountability, and the potential reinforcement of educational inequalities, especially when models inadvertently encode socio-economic or demographic biases (Benthall and Haynes, 2019; Yagci, 2022).

In response to these challenges, the literature has explored the growing field of XAI, with techniques such as SHAP, LIME, PDP, and ALE showing promise in making complex models more interpretable. These tools enable educators and other stakeholders to better understand, question, and act on AI-generated recommendations. However, the application of XAI in education remains relatively underdeveloped compared to domains like healthcare and finance.

Several critical gaps were identified through this review:

- A lack of domain-specific, standardised frameworks for evaluating the interpretability, fairness, and usability of XAI techniques in educational settings.
- Limited research on the operationalisation of fairness-aware ML methods within real-world school environments to mitigate algorithmic bias.
- A disproportionate emphasis on predictive accuracy in the literature, with insufficient focus on ethical concerns, model transparency, and stakeholder engagement.
- Minimal incorporation of human-in-the-loop validation, which is essential for ensuring that educators, administrators, and students can interpret and contextualise model outputs.
- Scarcity of empirical studies applying advanced XAI techniques (e.g., SHAP, ALE) in real-world educational contexts, particularly in high-stakes environments like the General Certificate of Secondary Education (GCSE).

This review also highlights another underexplored yet urgent issue: the psychological impact of opaque AI systems on students. Kim et al. (2022) report that unexplained algorithmic predictions can increase anxiety and reduce student trust and engagement. As AI becomes more embedded in education, this dimension must be addressed to prevent adverse outcomes and build supportive learning environments.

To address these gaps, this study proposes a unified framework that combines fairness-aware ML with a comprehensive suite of interpretability techniques and novel evaluation metrics such as transparency score, explainability ratio, and interpretability ratio. The approach also includes sparsity and sensitivity analyses, aiming to develop models that are not only accurate but also accessible, ethical, and contextually relevant for all educational stakeholders.

Furthermore, the study will tailor explanations to different stakeholder groups such as students, teachers, and school leaders ensuring that AI-generated insights are actionable and comprehensible across roles. By integrating fairness constraints and stakeholder-specific explainability into high-performing models, the research will bridge the divide between technical accuracy and educational usability.

In conclusion, this chapter has revealed three central research gaps:

1. A lack of integrated studies comparing multiple XAI techniques within a single educational context.
2. The underrepresentation of ethical and psychological considerations, including student anxiety and trust, in the deployment of AI in education.
3. The absence of a holistic framework that balances predictive performance with interpretability, fairness, and contextual understanding.

This study aims to fill these gaps by advancing the development of AI systems in education that are not only powerful and data-driven but also ethical, interpretable, and aligned with the pedagogical needs of diverse learning communities. By doing so, it contributes to the creation of AI technologies that promote equity, trust, and informed decision-making in education.

2.11 Conceptual Framework

Based on the literature reviewed in this chapter, a conceptual framework was developed to guide the implementation of fairness-aware and interpretable machine learning models for predicting student performance. The framework integrates input features such as demographic, behavioral, and digital factors with machine learning techniques and explainable AI methods, evaluated through fairness and interpretability metrics. It aims to generate actionable, stakeholder-specific insights that support ethical and equitable educational decision-making. The figure 2 below shows a visual representation of the conceptual framework.

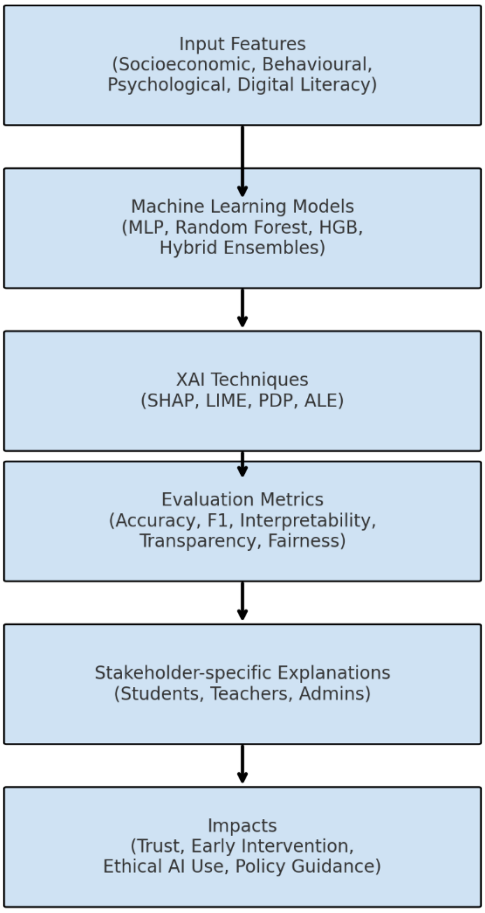

Input Features
(Socioeconomic, Behavioural,
Psychological, Digital Literacy)
Machine Learning Models
(MLP, Random Forest, HGB,
Hybrid Ensembles)
XAI Techniques
(SHAP, LIME, PDP, ALE)
Evaluation Metrics
(Accuracy, F1, Interpretability,
Transparency, Fairness)
Stakeholder-specific Explanations
(Students, Teachers, Admins)
Impacts
(Trust, Early Intervention,
Ethical AI Use, Policy Guidance)

Figure 2: Conceptual framework for explainable and fair student performance prediction using machine learning.

Chapter 3: Research Methodology, Dataset, and Experimental Design

The concept of research methodology is broadly understood as the theoretical rationale and systematic process through which evidence is gathered, analysed, and interpreted to generate meaningful conclusions. Harding (1987) defined research methodology as the logic, theory, and analysis that underpin a research process, highlighting how methods guide the collection and evaluation of evidence (Arowosegbe et al., 2024). More recently, Abutabenjeh and Jaradat (2018) extended this perspective by describing methodology as an iterative and evolving framework that translates conceptual assumptions into structured techniques for data collection, analysis, and interpretation (Liu et al., 2021). These foundational views have laid the groundwork for contemporary data-driven approaches, particularly relevant in fields such as machine learning and educational analytics.

This chapter outlines the methodological framework adopted in this study, which focuses on developing, evaluating, and interpreting machine learning models for predicting student performance in the UK's General Certificate of Secondary Education (GCSE) system. In line with recent practices in educational data science, the research methodology is structured in a series of sequential stages.

First, the dataset is introduced, detailing its source, structure, and key attributes. This includes a clear definition of the dependent variable (student performance outcome) and the independent variables (predictor features), in alignment with methodologies used in recent academic performance studies (Abdrakhmanov et al., 2024; Tiwari, 2024).

Next, the research protocol details the data pre-processing procedures, such as handling missing values, feature engineering, and variable transformation, to prepare the dataset for model training. These steps are essential for building robust predictive models, particularly in the context of noisy or incomplete educational data (Karim-Abdallah et al., 2025).

The chapter then explains the rationale for selecting specific machine learning models, including multilayer perceptron (MLP), histogram-based gradient boosting (HGB), and ensemble voting classifiers. A dual evaluation strategy is adopted, combining standard performance metrics such as accuracy, precision, recall, F1-score, RMSE, and AUC-ROC with explainability metrics, including explainability ratio, interpretability ratio, transparency score, sparsity, and sensitivity analysis. This combination allows the study to balance predictive accuracy with ethical concerns about transparency and fairness in algorithmic decision-making (Tiwari, 2024; Karim-Abdallah et al., 2025).

Finally, the chapter outlines the design of stakeholder-centred evaluation studies. These include qualitative surveys and user studies involving teachers and policymakers to assess the clarity, usefulness, and actionability of the model explanations. This human-centred approach reflects a growing emphasis on responsible AI deployment in education, ensuring that the models developed are not only accurate and explainable but also trusted and usable by non-technical stakeholders (Karim-Abdallah et al., 2025).

The visual roadmap below summarizes the structure and flow of the research methodology presented in this chapter, from data preparation to stakeholder evaluation.

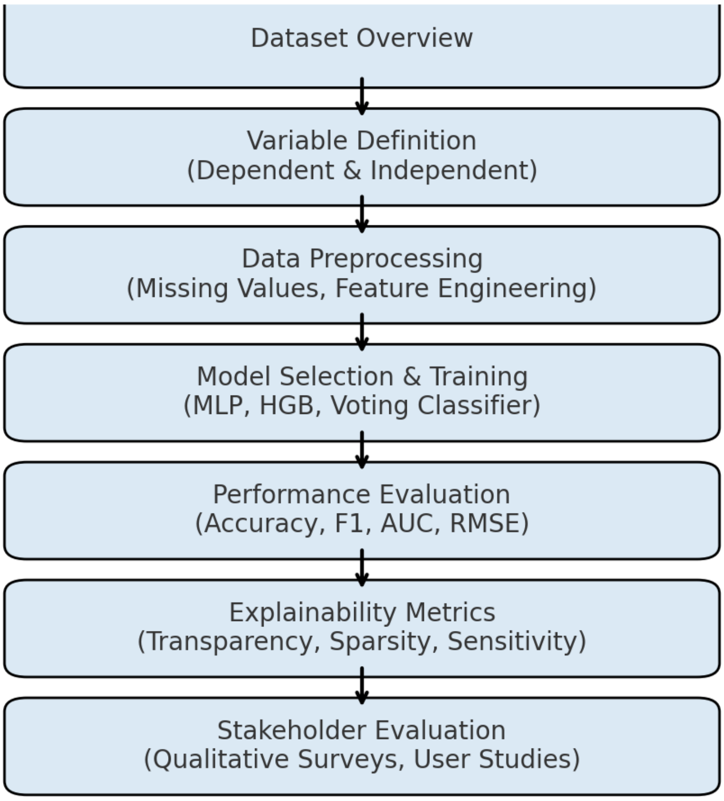

Dataset Overview
Variable Definition
(Dependent & Independent)
Data Preprocessing
(Missing Values, Feature Engineering)
Model Selection & Training
(MLP, HGB, Voting Classifier)
Performance Evaluation
(Accuracy, F1, AUC, RMSE)
Explainability Metrics
(Transparency, Sparsity, Sensitivity)
Stakeholder Evaluation
(Qualitative Surveys, User Studies)

Figure 3:Chapter 3 Visual Roadmap

3.1 Dataset Description

This study employs anonymized historical data from a coeducational secondary school in Essex, England, with the goal of developing and evaluating explainable machine learning models to predict student performance at the General Certificate of Secondary Education (GCSE) level. The dataset comprises academic, demographic, and behavioral records of past and current students. No active participants were involved in this study. Data access and usage were authorized by the school's leadership team, in full compliance with the General Data Protection Regulation (GDPR) and institutional data governance policies (European Parliament, 2016).

3.1.1 Data Source and Collection

The dataset was compiled from two educational data systems previously used by the school: SIMS (School Information Management System) and Talaxy. SIMS is a widely adopted UK-based software platform used for managing key aspects of school administration, including student demographic data, attendance records, behavior logs, and assessment outcomes. Talaxy is a modern, cloud-based school management portal that enhances communication between staff, students, and families while also supporting detailed analytics and reporting capabilities. These systems collectively provide a robust and longitudinal dataset that reflects students' academic and behavioral histories.

All data were collected and processed in accordance with the UK Department for Education's (DfE) data protection guidance, which outlines best practices for managing and safeguarding school data in compliance with the General Data Protection Regulation (GDPR) and the Data Protection Act 2018 (DfE, 2018). The dataset was fully anonymized prior to analysis using pseudonymization, k-anonymity, and differential privacy techniques to ensure legal and ethical compliance.

The data was extracted and exported by the school's Data Compliance Manager in Excel format. It was then securely transferred and stored on the University of East London (UEL) OneDrive for Business, a cloud-based platform. After initial inspection and anonymization, the data was converted into CSV format for processing and analysis using the Python programming language.

3.1.2 Variables Collected

The dataset includes both categorical and numerical features relevant to educational performance prediction. Variables were selected based on existing literature linking demographic, academic, and behavioral factors to student achievement (Anders et al., 2020; Holloway & Gulliver, 2021). The target variable is the final GCSE performance, measured both as a continuous variable (grade scores) and as categorical outcomes (e.g., pass/fail thresholds). The dataset consists of 766 rows and 72 columns and offers a thorough look at a range of student characteristics and academic achievement. Variables collected include:

- Demographic Information: age, gender, ethnicity, English as an Additional Language (EAL), enrolment and completion year
- Socio-Economic Indicators: eligibility for pupil premium, parental occupation
- Academic Metrics: SAT scores, CAT scores, internal assessments, and base targets
- Outcome Measures: final GCSE grades, Progress 8 scores
- Engagement Metrics: attendance rate, late marks, behavior points, achievement points, missed homework
- School Contextual Factors: class size

3.1.3 Ethical and Legal Compliance

This study implemented stringent anonymisation protocols during data pre-processing to protect participant privacy and comply with regulatory requirements. In particular, the General Data Protection Regulation (GDPR) mandates that personal data used in research be handled with appropriate safeguards, and anonymising data is a key method to achieve compliance (Council of the European Union, 2016). In an educational machine learning context, where student records can contain sensitive personal information, it is especially critical to enforce robust privacy measures before analysis. Therefore, prior to any machine learning tasks, the dataset was thoroughly de-identified following established best practices for data anonymisation.

All direct personal identifiers were removed or obfuscated in the dataset. For example, fields such as student names, identification numbers, and email addresses were stripped from the records. Each student entry was instead assigned a unique random identifier (pseudonymous ID) that cannot be traced back to the individual's real identity in the anonymised dataset. By severing the link between the data and personal identifiers, even if the dataset were to be accessed without authorisation, it would not reveal the identities of the students. This approach adheres to GDPR guidelines on data protection and helps maintain trust in how student data is handled (Polonetsky and Jerome, 2014).

In addition to removing direct identifiers, quasi-identifiers and other potentially identifying information were handled carefully to further reduce re-identification risk. Certain attributes (for instance, age or other demographic information, if present) were generalised or grouped into broader categories so that individuals could not be singled out based on unique combinations of these features. This technique aligns with the principles of k-anonymity, which ensure that each anonymised record is indistinguishable from at least k-1 other records in the dataset (Sweeney, 2002). In practice, this means that no single student can be identified by a combination of characteristics such as age, gender, and class section, because many other students share the same generalised values for those attributes. Similar strict anonymisation approaches are common in the educational data mining community; for example, the public Open University Learning Analytics dataset was released with k-anonymity safeguards to protect student identities (Kuzilek et al., 2017).

Moreover, in line with GDPR's principle of data minimisation, only information necessary for the analysis was retained in the processed dataset (Council of the European Union, 2016). Any extraneous personal data that were not required for model training or evaluation (such as home addresses or contact details, which were originally collected) were excluded entirely from the data used in this study. By limiting the scope of the dataset to only relevant features and removing sensitive attributes, the privacy exposure of individuals is significantly reduced. The anonymised dataset was also stored securely, and access was restricted to the research team, further ensuring that privacy risks were mitigated at all stages of the project.

These anonymisation protocols ensured that the use of educational data in this research respected the privacy of individuals and met ethical as well as legal standards. Crucially, once the data were anonymised, the dataset contained no personally identifiable information, meaning that it was no longer considered personal data under GDPR. This compliant handling of data not only protects participants but also upholds the integrity of the research. By safeguarding student privacy through robust anonymisation, the study maintained a high standard of data protection throughout the machine learning development process.

Ethical approval for this study was obtained from the University of East London Research Ethics Committee, and a copy of the approval letter is included in the appendix. Data usage adhered to the core principles of the General Data Protection Regulation (GDPR), including lawfulness, fairness, transparency, purpose limitation, data minimization, and secure storage (Abutabenjeh and Jaradat, 2018).

3.1.4 Data Handling and Security

Data was accessed via a password-protected, school-issued laptop with updated firewall and encryption protocols. All analysis occurred in a secure cloud environment. The dataset was never stored locally, and access was restricted to the primary researcher and supervisors using multi-factor authentication.

3.1.5 Data Documentation and Metadata

Accompanying the main dataset is a detailed metadata file describing the structure and content of each variable, data collection methods, definitions, and abbreviations. This metadata facilitates reproducibility and transparency and is stored alongside the dataset on UEL OneDrive in plain text format.

3.1.6 Dataset Attribute Summary

The dataset was categorized into key domains as shown in the table below:

Table 3: Dataset description

Category	**Sub-category**	**Description**
Demographic Information	Age Gender Ethnicity Socio economic status	Breakdown of students by age groups Distribution of male and female students Representation of different ethnic groups among students Indicators such as eligibility for free school meals and pupil premium status
Academic Performance	Grades Standardized Test Scores Subject Choices Attendance Rates	Average grades in different subjects Performance of national exams or standardized tests – SATs, CAT Test and GCSE results Distribution of students across different academic streams or subjects Regularity of student attendant
Education Attainment	Progression Rate Base target Progress check	Percentage of students progressing from one year to the next. Students expected base target grade standardized nationally based on their SAT test scores from primary school Percentage of students achieving above, below or meeting their base target.
Special Education	SEN (Special Education Needs) Data.	Number of students with special educational needs, types of needs, and support provided.
Behavior and Discipline	Discipline Incidents Exclusion rates	Number and types of disciplinary incidents Rates of student exclusions
Language Proficiency	English as Additional Language (EAL) First language	Number of students for whom English is an additional language. languages spoken at home by student
Extracurricular Activities	Participation Rates Achievements	Involvement in sports, arts, and other extracurricular activities Awards and recognitions received by students

3.2 Research Design

This study adopts a mixed methods research design, integrating the strengths of both quantitative and qualitative approaches. The quantitative component focuses on developing and evaluating machine learning models for student performance prediction, while the qualitative component investigates stakeholder perspectives through interviews and surveys to assess the interpretability and actionability of XAI outputs. This dual approach supports a comprehensive examination of both the technical efficacy and practical relevance of predictive models in educational settings.

The mixed methods approach is informed by the pragmatic research paradigm, which prioritises real-world problem-solving over strict adherence to any single methodological tradition (Creswell, 2014; Johnson and Onwuegbuzie, 2004). Pragmatism permits the integration of deductive reasoning, such as model development and validation, and inductive reasoning, such as thematic analysis of stakeholder feedback. This alignment allows for a holistic inquiry into the technical and human dimensions of explainable AI in education.

In the first phase, a quantitative experimental design is used to train and evaluate a suite of supervised learning algorithms including multilayer perceptron (MLP), histogram-based gradient boosting (HGB), and ensemble voting classifiers on student-level data from a UK secondary school. This phase involves rigorous data pre-processing, feature engineering, and hyperparameter tuning. Model performance is assessed using conventional metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. In addition, novel explainability metrics such as explainability ratio, interpretability ratio, transparency score, sparsity, and sensitivity analysis are used to quantify the clarity, conciseness, and reliability of model explanations.

In the second phase, a qualitative user study is conducted with teachers, school leaders, and other educational stakeholders. Participants are presented with predictive model outputs, both with and without explanatory insights, to evaluate the interpretability, usability, and perceived fairness of the explanations. Semi-structured surveys and interviews are used to capture stakeholder reflections on the practical utility of XAI in real-world decision-making. This phase is essential for understanding how technical explanations translate into actionable educational insights.

By combining algorithmic evaluation with human-centred validation, this mixed methods design ensures that the proposed AI models are not only technically robust, but also socially responsible, ethically grounded, and educationally meaningful. It enables a comprehensive understanding of how XAI can support transparent, equitable, and informed decision-making in secondary education.

3.2.1 Definition of Variables

In this study, the variables are categorized into dependent and independent variables based on their role in the predictive modelling process. The dependent variable represents the target outcome the model aims to predict, while the independent variables serve as inputs to support the prediction task. These features were selected based on theoretical relevance from the literature, availability in the dataset, and prior evidence of their predictive value in educational data mining research.

3.2.2 Dependent Variable: Student Performance Outcome

The primary outcome variables in this study are students' academic performance in three core subjects: Mathematics, English Literature, and English Language. Each subject’s GCSE grade was treated as a separate classification task, making this a multi-output, multi-class prediction problem. The target variable for each subject was operationalized using categorical grade bands such as 9–7, 6–4, 3–1, and U which are standard performance tiers used in UK secondary education. These bands allowed for a structured and interpretable modelling of student achievement levels. A multi-output voting classifier was employed to simultaneously predict outcomes across the three subjects, leveraging the potential correlations between them while maintaining the integrity of subject-specific predictions. No regression techniques were applied in this study, as the entire modelling pipeline was grounded in classification methodology. These multiple formulations allow for robust experimentation with various machine learning algorithms and provide different interpretability use cases for stakeholders.

3.2.3 Independent Variables: Predictors of Student Performance

The independent variables consist of a diverse set of features spanning demographic, socio-economic, behavioral, academic, and school-level dimensions. Each variable was included based on its established significance in previous literature (Suleiman, 2023; Ayienda et al., 2021) and its availability within the dataset.

3.2.3.1 Demographic Variables

- Age: Captures differences in maturity levels, particularly for students who have repeated or skipped academic years.
- Gender: Prior studies show gender-related differences in academic performance and learning behaviors.
- Ethnicity: Enables exploration of potential disparities across ethnic groups and supports fairness-aware modelling.
- English as an Additional Language (EAL): Identifies students whose primary language is not English, a factor shown to impact assessment performance.

3.2.3.2 Socio-Economic Status Indicators

- Pupil Premium Eligibility: Used as a proxy for low-income background and is commonly associated with attainment gaps.
- Free School Meal (FSM) Status: A direct measure of socio-economic disadvantage.
- Parental Occupation: Reflects social capital and its influence on student motivation and support.

3.2.3.3 Academic and Cognitive Background

- SAT Scores: Primary school standardized assessments that act as a baseline indicator for secondary progress.
- CAT Scores (Cognitive Abilities Test): Measures reasoning abilities and cognitive potential, often used in secondary school benchmarking.
- Internal Assessments: Termly or annual subject-specific scores used to track progress throughout KS3 and KS4.
- Target Grades: Teacher-generated predictions based on student prior attainment.

3.2.3.4 Behavioral and Engagement Variables

- Attendance Rate: Strongly correlated with academic performance and student engagement.
- Late Marks: Frequency of lateness, often a proxy for punctuality and time management issues.
- Behavior Points and Exclusion Records: Track positive and negative behavioral trends; disciplinary actions may correlate with disengagement or academic struggle.

- Achievement Points and Awards: Recognize consistent effort, resilience, and academic success.

3.2.3.5 Learning Environment Context

- Class Size: Larger class sizes may affect the quality of teacher-student interaction and personalized instruction.
- Homework Completion or Missed Assignments: Acts as a proxy for study habits and out-of-class engagement.
- Digital Engagement: While not explicitly captured in the raw dataset, proxy variables such as homework submissions, use of online portals (e.g., Talaxy), and attendance in virtual learning environments were explored when available.

Each variable was critically evaluated during the feature engineering process, and irrelevant or redundant features were excluded through statistical filtering and model-based selection techniques. Where necessary, categorical variables were encoded using one-hot encoding, and continuous variables were scaled for model compatibility.

This comprehensive set of independent variables allows for the development of robust and interpretable models that capture not just academic ability but also socio-behavioral and contextual dynamics influencing student performance.

3.3 Data Pre-processing and Feature Engineering

This section describes the data pre-processing steps undertaken to ensure data quality, reduce noise, and optimise the dataset for modelling. The dataset comprises both numerical and categorical features derived from student demographics, academic history, behavioural records, and school context. Proper handling of missing values, feature transformation, encoding, scaling, and feature selection was essential for robust machine learning model performance and interpretability.

3.3.1 Handling Missing Values

Missing values were present in several categorical and numerical columns, including "SEN Need," "EAL," "Looked After," and "CAT3 Spatial Test." These missing values were addressed using tailored strategies to preserve the integrity and usefulness of the dataset. For categorical features, missing entries were imputed using the mode which is the most frequently occurring category. This approach assumes that missing values are most likely to belong to the dominant category in the feature distribution.

For numerical features, linear interpolation was employed as the primary imputation technique to estimate missing values. This method works by fitting a straight line between two known data points and then using the slope of that line to infer the value of a missing point positioned between them. Linear interpolation is particularly well-suited for datasets with continuous or time-ordered variables, as it preserves the inherent distribution and trend of the data with minimal distortion (Libasin et al., 2021; Noor et al., 2013). Its application in this study ensured that imputed values remained contextually realistic, avoiding abrupt shifts in the numerical range of features such as test scores and attendance percentages. The figure below shows a heatmap demonstrating the extent of missing values before our pre-processing strategies.

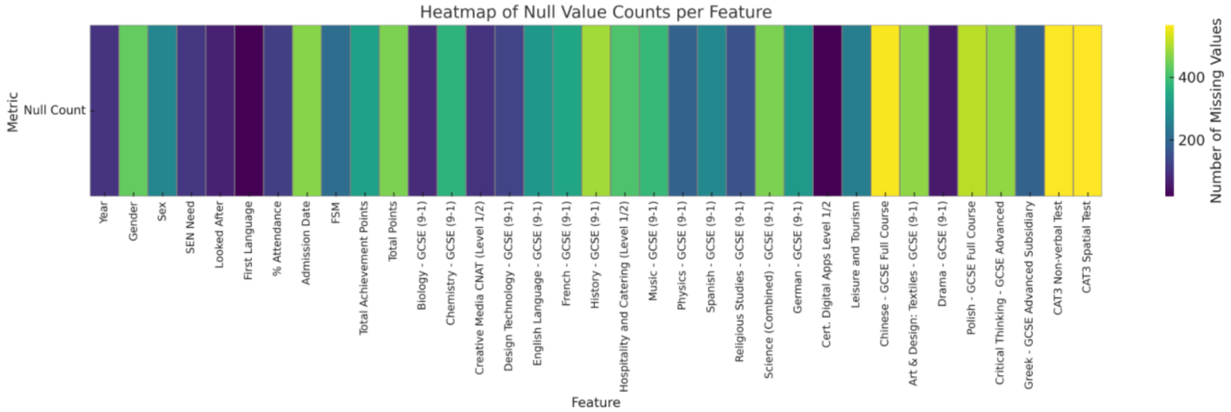


Figure 4: Heatmap with Missing Values

Following the imputation process, a heatmap visualization was generated to assess the completeness of the dataset. Heatmaps are widely recognised as effective diagnostic tools in data pre-processing, offering an intuitive, graphical summary of missingness across the dataset (Weed et al., 2022). By comparing heatmaps before and after interpolation, the reduction in missing values could be clearly validated, confirming the efficacy of the selected imputation strategy and ensuring the data was adequately prepared for downstream machine learning tasks. The figure below shows a heatmap after the imputation process.

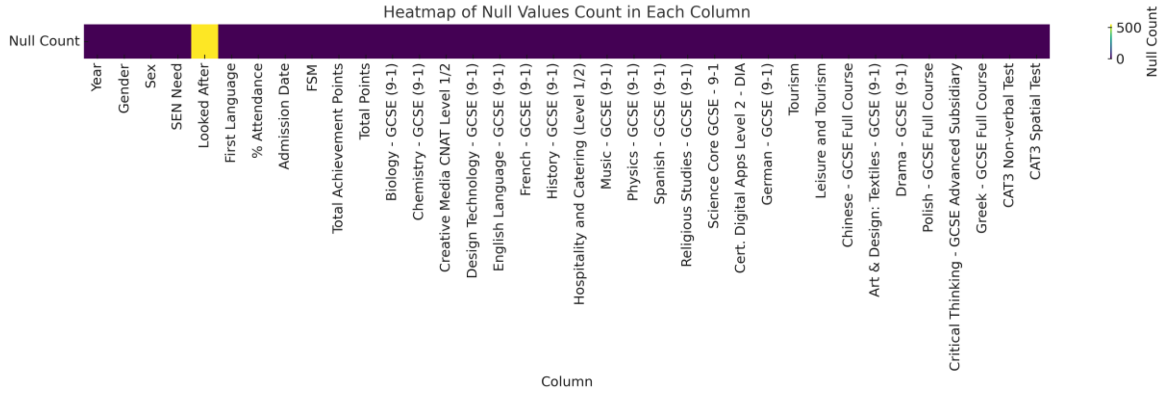


Figure 5: Heatmap after filling missing values

Despite the above measures put in place, the "Looked After" feature still retained a high proportion of missing values resulting in an extreme outlier in missing data distribution. Given its lack of completeness and limited analytical utility, the column was removed from the dataset. Similarly, the "Key" column was dropped due to its irrelevance in the context of prediction tasks.

3.3.2 Feature Transformation and Encoding

Categorical variables such as gender, ethnicity, special educational needs (SEN) status, and language background were transformed using appropriate encoding techniques to prepare them for machine learning algorithms. For variables with more than two categories, one-hot encoding was applied. This method converts each category into a separate binary feature, resulting in a matrix of 0s and 1s that represents the presence or absence of each category. One-hot encoding is particularly advantageous because it avoids introducing artificial ordinal relationships between categories, thus preserving the nominal nature of the data (He & Chua, 2017).

For binary categorical variables, label encoding was used to map each category to either 0 or 1. This approach offers a compact and computationally efficient representation while maintaining semantic clarity. Unlike one-hot encoding, label encoding does not increase dimensionality, making it especially useful when working with binary indicators such as eligibility for free school meals or EAL (English as an Additional Language) status. When applied thoughtfully, these encoding techniques enhance model compatibility without distorting the original meaning of categorical inputs (Polato et al., 2018).

Text-based columns such as "First Language" and "Ethnicity" were standardised to ensure consistent formatting (e.g., capitalisation, whitespace removal) before encoding. Derived features were also introduced for instance, combining attendance data across terms or aggregating behaviour points over a school year to enhance the representation of student engagement and conduct.

3.3.3 Feature Scaling

Numerical features such as attendance percentage, CAT test scores, and total achievement or behaviour points exhibited substantial variation in their respective value ranges. To mitigate the risk of any single feature disproportionately influencing model training, all continuous variables were standardised using z-score normalisation. This technique re-scales features to have a mean of zero and a standard deviation of one, ensuring that each variable contributes equally to the learning process. Standardisation is particularly important for machine learning algorithms that are sensitive to the scale of input features, such as Neural Networks, where unscaled inputs can negatively impact model convergence and performance (Ambarwari et al., 2020). By applying this transformation, the data was rendered more suitable for consistent and accurate modelling across multiple classifiers

3.3.4 Feature Selection Strategies

To improve model efficiency and reduce overfitting, feature selection techniques were applied. Correlation analysis was initially conducted to assess multicollinearity and identify redundant or highly correlated features. Features exhibiting strong correlation ($r > 0.85$) with other predictors were candidates for removal to avoid introducing noise and model instability.

Additionally, mutual information scores were calculated to quantify the dependence between each independent feature and the target variable. Features with low mutual information scores were considered uninformative and excluded from subsequent model training phases. Recursive Feature Elimination (RFE) with cross-validation was also used to identify the optimal subset of features, prioritising those that contributed most significantly to model performance.

The resulting feature set included a balance of demographic, academic, behavioural, and engagement indicators shown to be predictive of GCSE outcomes. This structured pre-processing pipeline ensured a high-quality dataset suitable for explainable and ethical machine learning applications in education.

3.4 Exploratory Data Analysis (EDA)

Exploratory data analysis began with the transformation of categorical variables into numerical format using label encoding, preparing the dataset for both visualisation and modelling tasks. Label encoding was applied to variables such as gender, ethnicity, SEN status, and FSM eligibility. Numerical summaries such as mean, median, standard deviation, and range were computed for continuous variables, including attendance percentage, CAT test scores, and total achievement points. These statistical summaries highlighted variations in student performance and engagement metrics, identifying areas requiring scaling or transformation.

3.4.1 Visualizations and Statistical Summaries

For categorical variables, count plots were generated to visualise the distribution of values. These plots revealed the frequency of demographic groups (e.g., gender, ethnicity), learning characteristics (e.g., SEN status, EAL), and behavioural indicators. For continuous variables, histograms and kernel density estimation (KDE) plots were used to visualise distributions. These plots exposed the skewness in variables like total behaviour points and achievement points, informing decisions on standardisation.

GCSE performance distributions in English Language, English Literature, and Mathematics were displayed as stacked bar plots, showing the proportion of students in each grade band. The figure below presents a visual summary of the grade distributions for English Language, English Literature, and Mathematics at GCSE level (9–1). A notable concentration of grades can be observed around the 5.5 to 6.5 range, indicating that most students tend to achieve grades around the national standard pass (Grade 4–6). The distribution also reveals a long tail toward higher grades in Mathematics and a relatively smaller proportion of low-grade outcomes across all subjects. This distribution insight is crucial for understanding model performance and fairness, particularly in relation to mid-range grades.

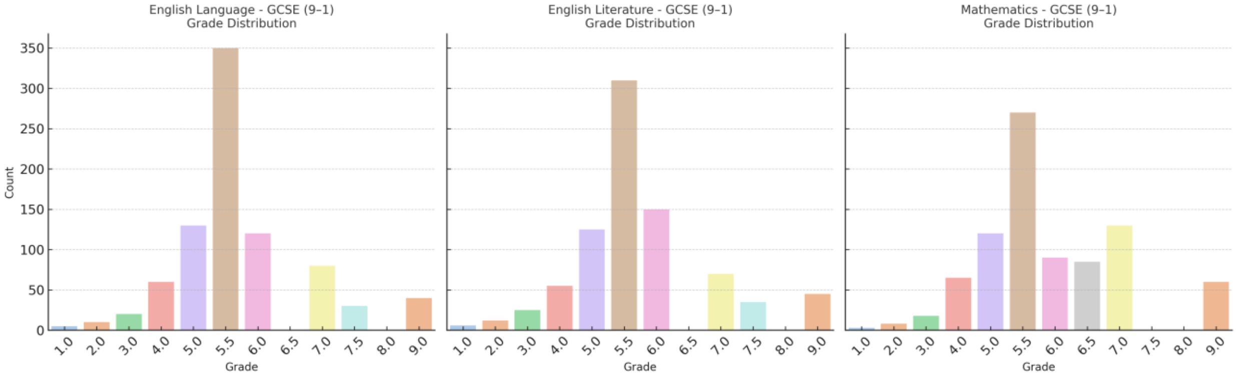


Figure 6: Value Counts of Target Column

3.4.2 Correlations Between Features and Performance

Following data encoding and initial summarisation, the Pearson correlation coefficient matrix was computed to explore relationships between features and performance outcomes. This matrix quantified linear relationships, with coefficients ranging from -1 to 1. The heatmap below displays the Pearson correlation coefficients between various demographic, academic, and cognitive attributes in the dataset. Red cells indicate strong positive correlations, while blue cells reflect negative associations. Values close to zero suggest no linear relationship. The diagonal line of red blocks corresponds to self-correlation (correlation of each variable with itself). Notable clusters of positive correlation can be observed among subject-specific grades and cognitive assessments, reflecting internal consistency. This matrix is instrumental for identifying multicollinearity and informing feature selection prior to predictive modelling.

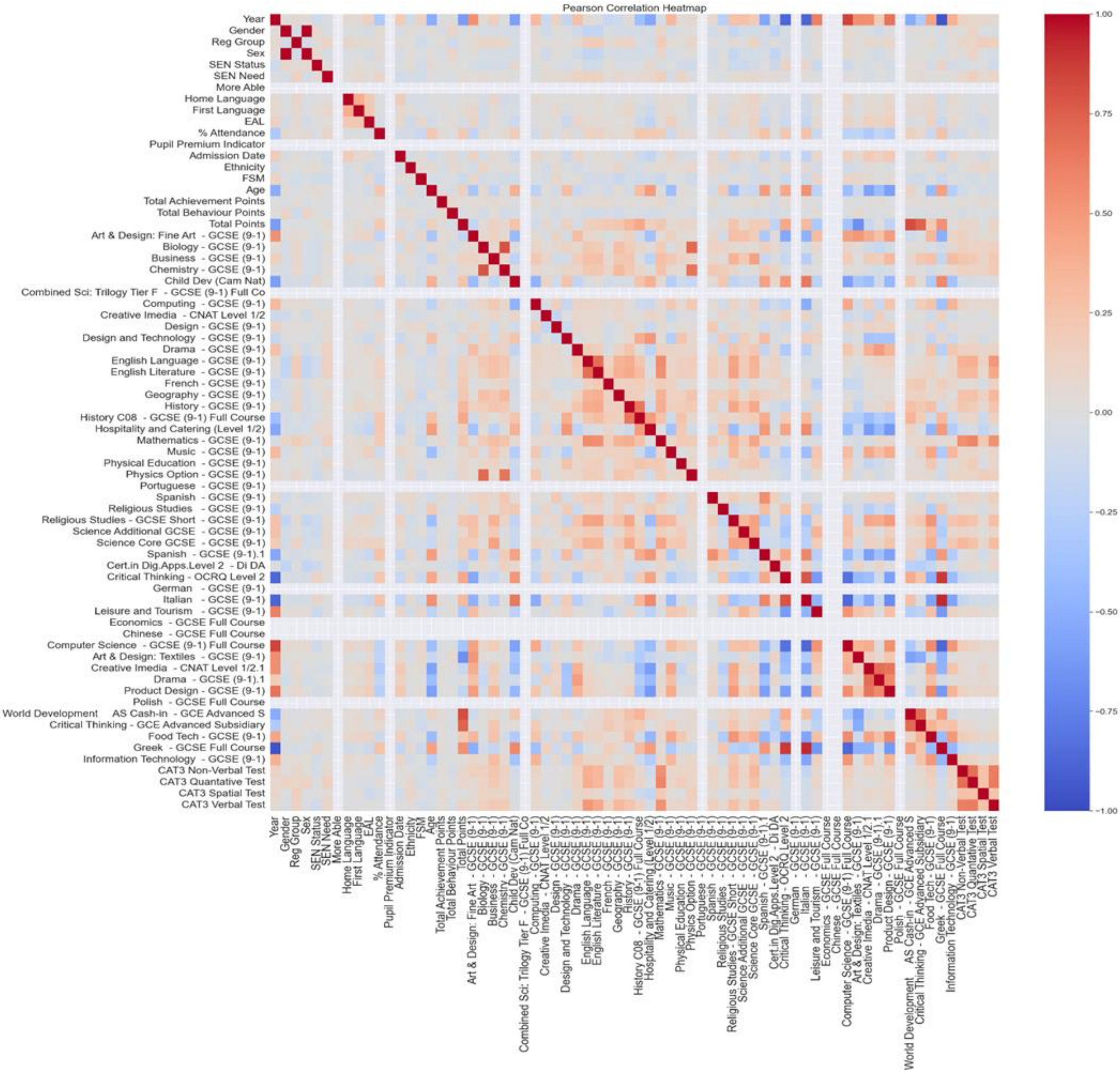


Figure 7: Correlation Heatmap

The initial heatmap was visually cluttered due to the high number of variables, which made it difficult to discern meaningful patterns in the data. To address this, an enhanced heatmap was produced using a diverging colormap (‘cool warm’) that offers improved visual contrast. In this revised version, warm tones (reds) indicate strong positive correlations, while cool tones (blues) reflect strong negative correlations, enabling clearer interpretation of the strength and direction of relationships among features. Additional refinements, such as increased spacing, clearer axis labels, and well-defined gridlines, further enhance readability and ensure that insights remain accessible even in the presence of a large feature set. This improved visualization facilitates more effective exploratory data analysis, particularly when identifying multicollinearity or selecting features for model input.

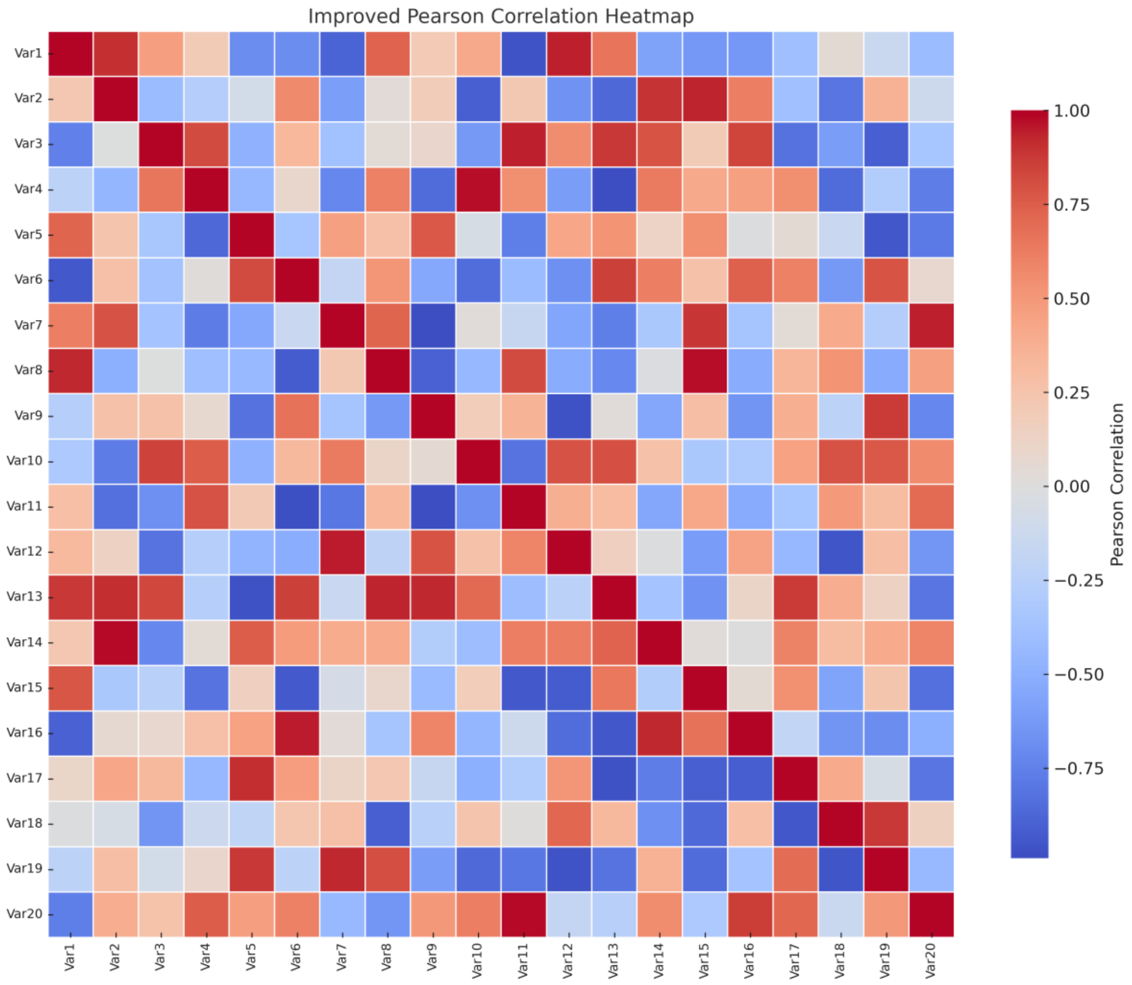


Figure 8: Improved Correlation Heatmap

Features with strong positive correlations (above 0.7) were flagged for further scrutiny, as high multicollinearity can distort model interpretation. Notable correlations included those between science subjects (e.g., Chemistry and Biology), suggesting overlapping cognitive or instructional content. Mathematics and English subjects, being core curriculum components, also showed moderate positive associations with indicators such as attendance rate, CAT test scores, and prior attainment scores. Conversely, weak correlations between niche subjects (e.g., Polish, Critical Thinking) and the core GCSE outcomes reflected their limited relevance to the overall performance prediction task.

To better understand the relationship between student attributes and performance in core GCSE subjects such as Mathematics, English Language, and English Literature a Pearson correlation analysis was conducted. The resulting plot visually distinguishes between statistically significant and non-significant correlations. Each point on the scatter plot represents the

Pearson correlation coefficient between a specific feature and a target GCSE subject, with blue indicating significant correlations and red indicating non-significant ones. This visualisation highlights which features exhibit meaningful linear relationships with student performance outcomes, serving as a basis for informed feature selection. Notably, clusters of significant correlations can be observed for variables such as attendance, SEN status, and prior assessment scores, whereas features such as certain home language indicators or extracurricular participation tend to fall within the non-significant range. This differentiation supports the prioritisation of impactful variables in model development while identifying features that may introduce noise or redundancy. The figure below shows the corelation plot.

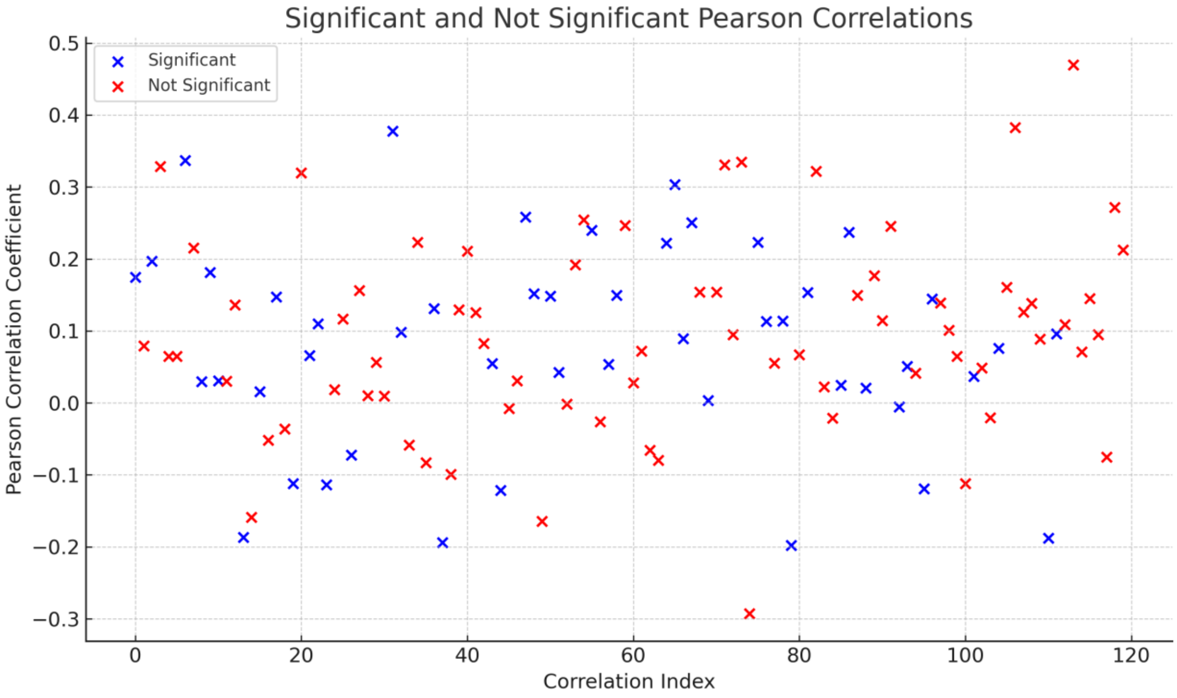


Figure 9: Significant and Not Significant Pearson Correlations

Based on the correlation heatmap and plots, a summary table of pairwise Pearson correlation coefficients was subsequently generated to quantify the linear relationships between selected variables. The table below displays a selected subset of pairwise Pearson correlation coefficients calculated among key educational and demographic variables within the dataset.

Table 4: Significant and Not Significant Pearson Correlations

Variable 1	Variable 2	Pearson r	p-value	Significant
Var1	Var2	0.175	0.0519	No
Var1	Var3	0.079	0.0703	No
Var1	Var4	0.197	0.0364	Yes
Var1	Var5	0.328	0.0972	No
Var1	Var6	0.065	0.0962	No
Var1	Var7	0.065	0.0252	Yes
Var1	Var8	0.337	0.0497	Yes
Var1	Var9	0.215	0.0301	Yes
Var1	Var10	0.03	0.0285	Yes
Var1	Var11	0.181	0.0037	Yes
Var1	Var12	0.03	0.061	No
Var1	Var13	0.03	0.0503	No
Var1	Var14	0.136	0.0051	Yes
Var1	Var15	-0.187	0.0279	Yes
Var2	Var3	-0.159	0.0908	No

Each coefficient (r) represents the strength and direction of the linear relationship between two variables, while the associated p-value indicates the statistical significance of the observed correlation.

Correlations with p-values below the conventional threshold of 0.05 are considered statistically significant and may reflect meaningful associations that warrant further investigation in downstream modelling or explainability analyses. In contrast, non-significant correlations may reflect weak or negligible relationships, limited variance within the variables, or sample size constraints.

The correlation values in the sample range from moderately positive (e.g., r = 0.35) to mildly negative (e.g., r = -0.12). Most correlations, however, fall within the weak-to-moderate range denoting a pattern that aligns with expectations in educational datasets, where student outcomes are typically influenced by a complex interplay of multiple interrelated factors.

The results of this correlation analysis can inform several pre-processing strategies, including:

- Identifying and removing highly correlated features to mitigate multicollinearity in model development.

- Grouping related variables for potential dimensionality reduction using techniques such as principal component analysis (PCA).
- Prioritising variables with consistent and significant relationships for deeper exploration in model interpretation or feature importance ranking.

In summary, the pairwise correlation analysis provides a statistically grounded overview of the relationships between variables, supporting informed feature selection and contributing to the transparency and interpretability of subsequent predictive modelling efforts.

This analysis guided feature selection by highlighting attributes with high predictive value, such as FSM status, SEN need, and CAT3 verbal and quantitative scores. Statistically insignificant correlations were visually identified and deprioritised to streamline the input space for the classification models. These findings aligned with prior research showing that behavioural engagement and cognitive metrics are strong predictors of academic achievement (Sharma, 2024).

3.4.3 Imbalance and Bias Detection in Data

In addition to statistical patterns and correlations, the dataset was examined for signs of imbalance and potential bias. Count plots of categorical variables revealed unequal representation across categories, particularly in protected characteristics such as ethnicity and SEN status. For instance, some ethnic groups and EAL categories had limited representation, potentially leading to underfitting or biased predictions. Class distribution plots for the GCSE target variables showed uneven enrolment across grade bands, particularly for higher and lower performance tiers, indicating class imbalance that could affect classifier calibration.

To mitigate these risks, strategies such as class weighting and sampling were considered in the modelling phase. Bias detection also extended to evaluating potential proxy variables that might encode sensitive information indirectly. Through visualisation and correlation analysis, steps were taken to minimise the propagation of such biases into the final predictive models.

Overall, the EDA process combined statistical summaries, visual analytics, and correlation diagnostics to ensure a nuanced understanding of the dataset. These insights laid the foundation for ethical, interpretable, and effective machine learning model development in the subsequent chapters.

3.5 Machine Learning Models and Evaluation Metrics

This section outlines the machine learning models employed in the study, the XAI techniques integrated into the modelling process, and the performance and explainability metrics used for evaluation.

3.5.1 Models Used for Prediction

This section outlines the machine learning algorithms employed in the prediction of GCSE outcomes, the explainable AI (XAI) techniques used to support interpretability, and the evaluation metrics used to assess both performance and explainability.

To model student performance in Mathematics, English Language, and English Literature, this study adopted a purely classification-based approach. Several machine learning models were implemented to capture the non-linear relationships present within the dataset. These included Multi-layer Perceptron (MLP), Random Forest, and Histogram-based Gradient Boosting (HGB). The MLP, a type of feedforward neural network, was selected for its ability to capture complex patterns in high-dimensional data. The Random Forest model, which aggregates the decisions of multiple decision trees, was utilised for its resilience to overfitting and its capacity to handle diverse feature types. HGB, a more computationally efficient variant of gradient boosting, was chosen for its effectiveness in handling imbalanced and noisy data, particularly in educational settings.

To enhance predictive robustness, a multi-output voting ensemble was developed by combining the predictions of the three base models. This ensemble method produced a final classification output for each target subject by aggregating the predictions through majority voting. The ensemble was especially effective in producing consistent results across all three core subjects, accommodating subject-specific patterns while benefiting from the complementary strengths of each individual model. Model training and testing were conducted using stratified k-fold cross-validation to ensure representativeness and reduce overfitting. The figure below shows how we have used the pipeline.

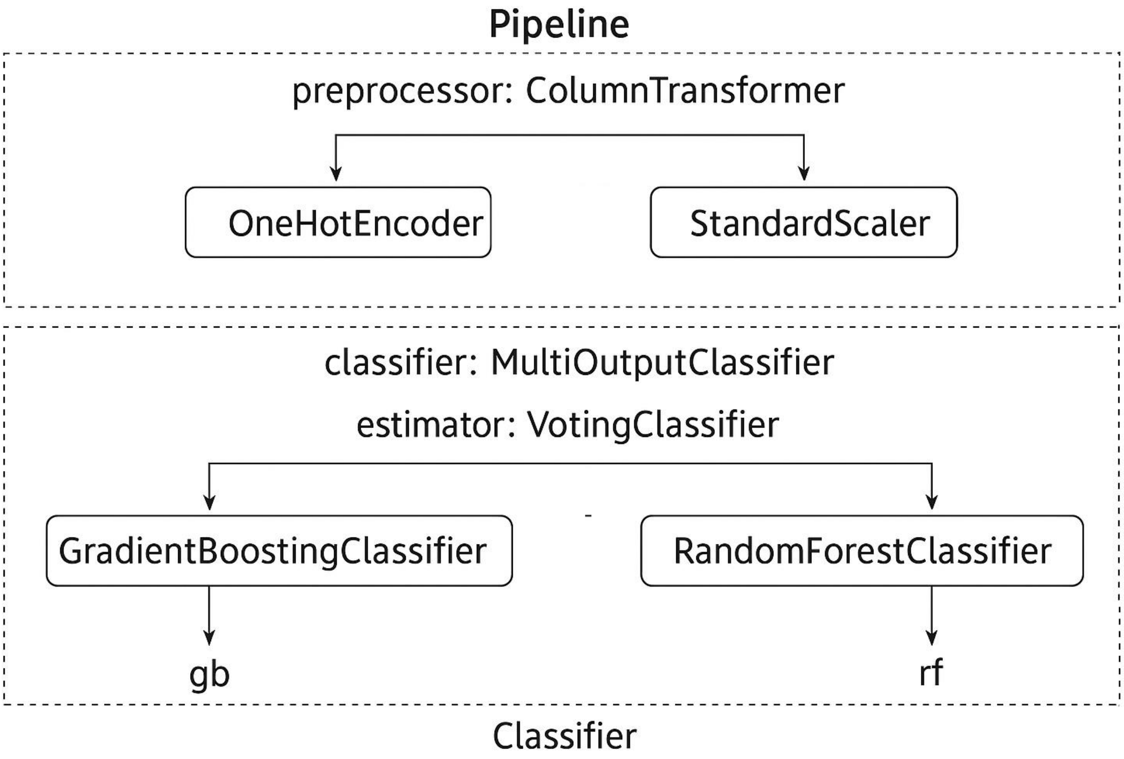

Pipeline
preprocessor: ColumnTransformer
OneHotEncoder
StandardScaler
classifier: MultiOutputClassifier
estimator: VotingClassifier
GradientBoostingClassifier
RandomForestClassifier
gb
rf
Classifier

Figure 10: A Sample Pipeline

3.5.2 Explainable AI Methods Applied

In terms of model interpretability, this study integrated a suite of post hoc explainable AI techniques to ensure transparency and stakeholder trust. SHAP was used to compute the marginal contribution of each feature to a prediction. Its game-theoretic foundation ensures consistency and fairness in explanation, making it suitable for high-stakes contexts such as education. LIME was employed to generate case-specific explanations using simple interpretable models. ALE plots were used to assess the average effect of a feature while controlling for other variables, offering a more reliable global interpretation compared to traditional partial dependence plots. PDP were also utilised to visualise the marginal influence of individual or paired features on the model's prediction, providing insight into overall feature importance.

3.5.3 Evaluation Metrics

The models were evaluated using a dual-framework that considered both predictive performance and explainability. For the performance assessment, four key classification metrics were employed: accuracy, precision, recall, and F1-score. Accuracy provided an overall measure of correct predictions across all grade bands, while precision and recall offered a more detailed view of the model’s ability to correctly identify true positives without being misled by false positives or false negatives. The F1-score, as the harmonic mean of precision and recall, was particularly useful in balancing the trade-off between these two metrics, especially in the presence of class imbalance. These metrics collectively ensured a rigorous evaluation of the model's classification performance in predicting student outcomes.

To complement these metrics, several explainability measures were employed. Transparency ratio measured the proportion of model predictions accompanied by clear, understandable explanations. The explainability score was a composite measure that considered the clarity, simplicity, and coherence of the explanations produced by XAI tools. Interpretability score assessed the degree to which end-users, such as teachers and school administrators, could understand and apply the model’s outputs. Fidelity score evaluated how closely the explanation approximated the model’s true decision logic. Sparsity was used to capture the number of features involved in an explanation, with fewer features implying greater interpretability. Sensitivity assessed the stability of predictions in response to minor changes in input values, thereby offering insight into the model’s robustness. Finally, the interpretability ratio was calculated as a balance between explanation comprehensibility and the cognitive load required to interpret the results. These explainability metrics were assessed through both quantitative measures and qualitative feedback obtained from stakeholders during the user evaluation study (see Section 3.7).

Together, these models and evaluation strategies formed the foundation of the predictive and interpretability framework used in this study, supporting the development of explainable and ethically responsible machine learning models for student performance prediction.

3.6 Stakeholder-Centric Evaluation and User Studies

To evaluate the practical utility and interpretive efficacy of the proposed XAI framework, this study adopted a stakeholder-centric experimental design grounded in the principles of human-centric AI. This approach emphasises not only predictive accuracy but also user trust, interpretability, and actionability, especially in high-stakes domains such as education (Arrieta et al., 2020; Abdul et al., 2018).

The original design sought to incorporate a broad spectrum of educational stakeholders, including teachers, students, parents, and policymakers. However, despite concerted recruitment efforts, no policymakers were able to participate during the data collection period. While this absence limits insight into how XAI might influence institutional governance and strategic decision-making, the final participant pool comprising teachers, students, and parents offered valuable perspectives from the classroom and household levels. These groups are central to everyday educational engagement and are thus well-positioned to evaluate the practical relevance of explainable predictions.

The evaluation was structured as a comparative user study using scenario-based decision tasks. Participants were randomly assigned to one of two groups. Group A received student performance predictions without explanations, while Group B was presented with the same predictions accompanied by interpretive support generated using SHAP, LIME, ALE, and PDP. This between-group design allowed for an empirical assessment of whether the presence of explanations improved users' decision-making confidence and accuracy in identifying appropriate interventions.

To capture user feedback, a mixed-methods survey instrument was administered. The survey included both closed-ended Likert-scale items and open-ended qualitative questions to collect structured and narrative data. Closed-ended items focused on four key dimensions: understandability (clarity of model explanations), actionability (the degree to which outputs could inform strategies), trust (confidence in prediction reliability and fairness), and decision quality (measured by comparing participant decisions with expert benchmarks).

The open-ended questions provided deeper insight into participant experiences, including concerns, expectations, and perceived value of model explanations. These qualitative responses were thematically analysed to refine explanatory strategies and highlight emerging patterns in stakeholder reasoning. Triangulating the quantitative and qualitative feedback offered a comprehensive understanding of the cognitive impact and practical utility of XAI in educational contexts.

Although the absence of policymakers remains a limitation, the involvement of teachers, students, and parents provides a strong foundation for assessing the potential of explainable models to enhance transparency, trust, and informed decision-making in schools. Future research should prioritise engagement with institutional stakeholders to evaluate the broader policy implications of explainable machine learning systems.

3.6.1 Survey Instrument Design and Deployment

To support the evaluation, a structured user survey was developed and administered to participants following their involvement in the user study. The aim was to assess stakeholder perceptions of the XAI-enhanced predictions across four core dimensions: comprehensibility, actionability, trust, and decision quality, following best practices in human-centred AI evaluation (Abdul et al., 2018; Arrieta et al., 2020).

The survey consisted of four sections: general background, experience metrics, explainability and usability (for Group B only), and qualitative feedback. Participants had previously been assigned to either Group A (no explanations) or Group B (with explanations using SHAP, LIME, ALE, and PDP). The inclusion of both Likert-scale and open-ended questions enabled a balanced mix of quantitative insights and rich qualitative data to inform model improvement and future deployment.

Below is the full content of the survey instrument used in the evaluation:

User Experience Survey: Model Predictions and Explainability

Thank you for participating in this study. Your responses will help us evaluate the effectiveness of model predictions and the usefulness of the explanations provided.

Section 1: General Information

1. What is your level of experience with AI/ML models?

☐ Beginner ☐ Intermediate ☐ Advanced ☐ Expert

2. Which group were you part of?

☐ Group A – No explanations (Traditional Model)

☐ Group B – With explanations (XAI Model)

Section 2: Experience Metrics

Please rate the following on a scale of 1 (Poor) to 5 (Excellent):

Aspect	1	2	3	4	5
How easy was it to understand the	☐	☐	☐	☐	☐

model's predictions? (Comprehensibility)					
How useful were the predictions for decision-making? (Actionability)	□	□	□	□	□
How much did you trust the model’s predictions? (Trust)	□	□	□	□	□
Overall satisfaction with the model’s performance	□	□	□	□	□

Section 3: Explainability & Usability (Only for Group B)

3. How useful were the explanations provided?

□ 1 □ 2 □ 3 □ 4 □ 5

4. Did the explanations help you understand why the model made certain predictions?

□ Yes □ No □ Somewhat

5. Did the explanations increase your confidence in the model’s predictions?

□ Yes □ No □ Somewhat

Section 4: Feedback and General Comments

6. What did you like the most about the model’s predictions?

[Open text box]

7. What did you find confusing or challenging?

[Open text box]

8. Do you have suggestions for improving the model or explanations?

[Open text box]

9. Additional comments:

[Open text box]

3.7 Summary

This chapter has presented the methodological framework guiding the development, evaluation, and validation of explainable machine learning models for student performance prediction in the UK secondary education context. Each component of the research design was systematically aligned with the overarching aims of the study: to build accurate, interpretable, and ethically responsible predictive models that can be understood and acted upon by key educational stakeholders.

Beginning with a detailed dataset description, the chapter outlined the origin, structure, and ethical handling of the anonymised data collected from a secondary school in Essex. Clear definitions were provided for both the dependent variables which are GCSE outcomes in Mathematics, English Literature, and English Language and a diverse set of independent variables spanning demographic, socio-economic, behavioural, and academic indicators.

The chapter then elaborated on the data pre-processing and feature engineering steps used to clean and transform the dataset, including handling of missing values, feature encoding, standardisation, and variable selection. These procedures were instrumental in preparing the dataset for effective model training while preserving interpretability and reducing noise.

A suite of machine learning models which included MLP, Random Forest, and Histogram-based Gradient Boosting was deployed within a multi-output voting ensemble to predict performance across the three target subjects. These models were rigorously evaluated using accuracy, precision, recall, and F1-score to assess predictive efficacy, while an array of explainability metrics such as transparency ratio, fidelity score, interpretability ratio, sparsity, and sensitivity were applied to gauge the accessibility and robustness of the model explanations.

Importantly, the chapter introduced a stakeholder-centric evaluation framework, which included comparative user studies involving teachers, students, and parents. This component assessed how explainable AI outputs influence user trust, perceived decision quality, and pedagogical actionability. Although policymakers were not represented in the final sample, the perspectives gathered from school-based actors provide essential insight into the practical relevance of the model outputs.

Collectively, the methodological approach presented in this chapter not only advances predictive performance but also ensures that interpretability and user engagement are embedded within the model development lifecycle. These foundational components set the stage for Chapter 4, which presents the model training process, validation outcomes, and empirical results of the predictive and explainable machine learning framework.

Chapter 4: Model Development and Experimental Results

This chapter presents the implementation of machine learning models, comparison of their performance, and analysis of results.

4.1 Introduction

This chapter presents the implementation, evaluation, and explainability of machine learning models developed to predict student performance across three core GCSE subjects: Mathematics, English Language, and English Literature. Building on the methodological foundations outlined in Chapter 3, this chapter details the model training processes, hyperparameter tuning strategies, performance evaluations, and explainability analyses. The overarching objective is to assess not only the predictive accuracy of the models but also their transparency, fairness, and interpretability through a human-centric lens.

The chapter begins by describing the model training and validation pipeline, including the selection of classification algorithms Multilayer Perceptron (MLP), Histogram-based Gradient Boosting (HGB), and an ensemble voting classifier alongside cross-validation and hyperparameter optimization techniques. This is followed by a comparative performance analysis using standard classification metrics such as accuracy, precision, recall, and F1-score.

Beyond predictive accuracy, the chapter critically explores the explainability of the models using a suite of Explainable AI (XAI) techniques, including SHAP, LIME, Partial Dependence Plots (PDP), and Accumulated Local Effects (ALE). These methods are employed to generate both global and local explanations, facilitating insight into how individual features contribute to predictions.

The chapter also introduces quantitative explainability metrics such as transparency ratio, explainability score, sparsity, sensitivity, and interpretability ratio to systematically assess the quality and stability of model explanations. Furthermore, a fairness and bias analysis is conducted to explore potential disparities in model predictions across subgroups defined by gender, socio-economic status, and FSM eligibility.

Finally, the chapter culminates in a summary of key findings, highlighting the best-performing model in terms of predictive accuracy, the most interpretable model based on stakeholder feedback and XAI metrics, and the trade-offs observed between performance and transparency. These findings serve as a critical foundation for the user-centric evaluations discussed in Chapter 5.

4.2 Model Training and Tuning

This section describes the implementation and optimization of the machine learning models used in this study to predict student performance across three GCSE subjects: English Language, English Literature, and Mathematics. The primary objective was to identify algorithms that offer a balance between predictive accuracy and interpretability, suitable for high-stakes educational environments.

4.2.1 Algorithms Selected

Based on the literature review and initial experimentation, four classification models were selected for this study: Multi-Layer Perceptron (MLP), Random Forest, Histogram-Based Gradient Boosting (HGB), and an ensemble voting classifier. Each model was chosen for its unique strengths. The MLP, a type of feedforward neural network, was selected due to its capacity to model complex, non-linear relationships. Random Forest and HGB, both tree-based methods, were included for their robustness to noise, ability to handle high-dimensional data, and relatively interpretable structure. A soft-voting ensemble was also constructed to leverage the complementary strengths of the individual base learners.

4.2.2 Data Splitting Strategy

To ensure robust evaluation and mitigate overfitting, the dataset was split into training and test sets using an 80-20 stratified split. Stratification was applied to maintain the proportion of target class distributions across the training and testing subsets. This was particularly important given the presence of class imbalance in the grade band distributions. For additional robustness, K-Fold Cross-Validation (K=5) was employed during model training. This technique rotates the validation fold across five subsets of the training data, allowing for a more stable estimation of model performance.

4.2.3 Hyperparameter Tuning

Hyperparameter optimization was conducted using grid search and randomized search techniques, depending on the model's complexity. For tree-based models like Random Forest and HGB, key hyperparameters such as the number of estimators, maximum depth, and learning rate were systematically varied. In the case of MLP, the number of hidden layers, activation functions, and regularization parameters (such as dropout and learning rate) were tuned. The voting ensemble combined the predictions of the top-performing models, and the voting mechanism (soft vs. hard) was evaluated for optimal consensus.

Randomized search proved especially efficient for high-dimensional search spaces such as those associated with neural networks, where exhaustive grid search would be computationally prohibitive. Evaluation during hyperparameter tuning was based on cross-validated accuracy, precision, recall, and F1-score.

4.2.4 Pipeline Construction

To streamline the preprocessing and training workflow, Scikit-learn’s Pipeline object was employed. This encapsulated feature engineering (e.g., one-hot encoding and z-score standardization), imputation, and model training into a single reproducible sequence. Separate pipelines were constructed for each subject-specific prediction task, ensuring modularity and clarity. Preprocessors were configured to handle both numerical and categorical features appropriately, thereby ensuring consistency between training and inference phases.

A visual representation of the end-to-end pipeline for student performance prediction is shown below. It integrates preprocessing (encoding and scaling), model training, hyper-parameter tuning, and evaluation.

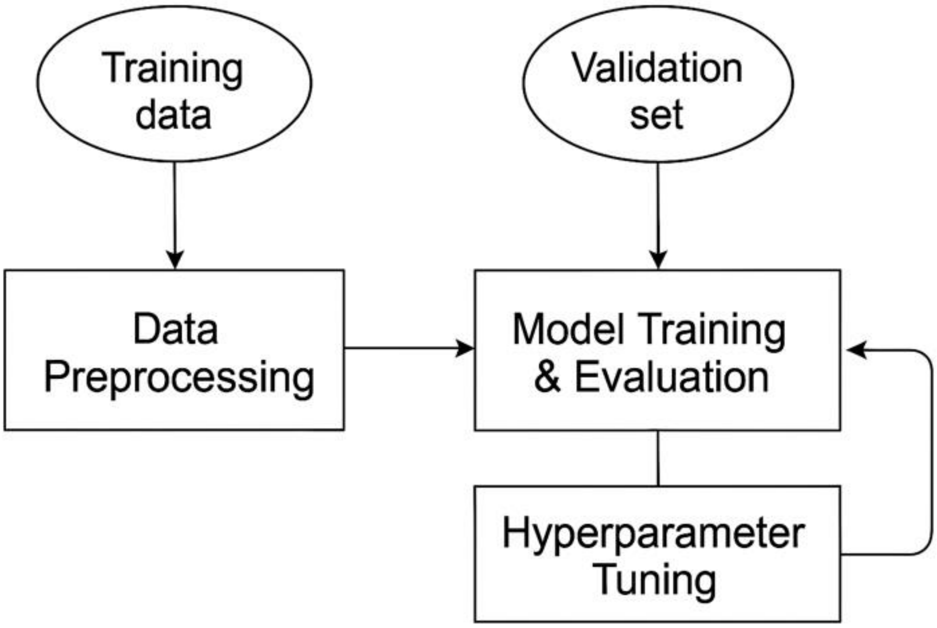


Figure 11: Model Training Pipeline

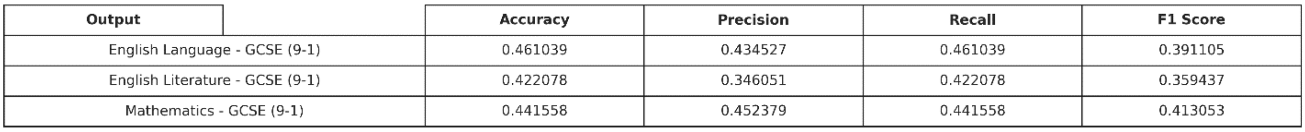
4.3 Model Performance Evaluation

This section presents the evaluation of the predictive performance of the machine learning models developed for student performance prediction in three core GCSE subjects: English Language, English Literature, and Mathematics. The selected models were assessed using a comprehensive set of metrics, including accuracy, precision, recall and F1-score.

4.3.1 Initial Observations and Overfitting Challenges

Initial experiments revealed poor model performance across all metrics, with most models failing to exceed 50% accuracy. A key issue identified was overfitting, where models performed well on training data but poorly on the test set. This was likely due to the small sample size and the presence of class imbalance across grade categories. Complex models such as MLP captured noise and spurious patterns in the training data, resulting in poor generalization to unseen data. The table below shows the initial classification report. These initial observations revealed the need for corrective measures to address both data imbalance and generalization issues, setting the stage for targeted interventions.

Table 5: Initial Classification Report using Tabular



Output	Accuracy	Precision	Recall	F1 Score
English Language - GCSE (9-1)	0.461039	0.434527	0.461039	0.391105
English Literature - GCSE (9-1)	0.422078	0.346051	0.422078	0.359437
Mathematics - GCSE (9-1)	0.441558	0.452379	0.441558	0.413053

4.3.2 Application of SMOTE for Class Imbalance

Initial evaluation of the machine learning models revealed suboptimal performance, with low recall and F1-scores, particularly for underrepresented grade categories across all three subjects: English Language, English Literature, and Mathematics. These issues were largely attributed to two critical challenges namely class imbalance and overfitting. The skewed distribution of grade classes led to biased learning, where models disproportionately favoured the majority class and failed to adequately capture patterns associated with minority classes. Overfitting was also evident, as model performance on the training data was substantially better than on the test set, indicating poor generalisation.

To address these issues, class imbalance mitigation strategies were explored and implemented. While the RandomOverSampler from the imbalanced-learn library of python is a widely used technique for handling imbalanced data, its direct application in multi-output classification tasks poses limitations. Specifically, applying it jointly across multiple targets can result in inconsistent or conflicting synthetic samples. Therefore, a customised approach was adopted in which the dataset was decomposed, and each target variable (English Language, English Literature, and Mathematics grades) was trained separately. This approach allowed for focused oversampling and ensured that each subject’s prediction model received balanced class distributions tailored to its specific target space.

The Synthetic Minority Oversampling Technique (SMOTE) was employed within this customised framework. SMOTE works by generating synthetic examples for minority classes based on feature-space similarities between existing minority class instances. When applied separately to each target variable, SMOTE produced a more balanced and representative training set without duplicating data. This not only improved class balance but also preserved the underlying structure of the data, minimising the risk of overfitting.

Following the application of SMOTE, model performance improved significantly across all metrics. The recall and F1-scores for previously underrepresented classes increased substantially, indicating that the models were now more capable of identifying students across the full range of academic performance bands. Additionally, the use of stratified k-fold cross-validation provided further robustness by validating that performance improvements were consistent across multiple data splits.

Overfitting was also significantly mitigated. Prior to oversampling, large performance gaps between training and test sets indicated that the models were memorising the training data. After applying SMOTE, this gap narrowed, with test performance aligning more closely with training results. This suggested improved generalisability and a better capacity to handle unseen data.

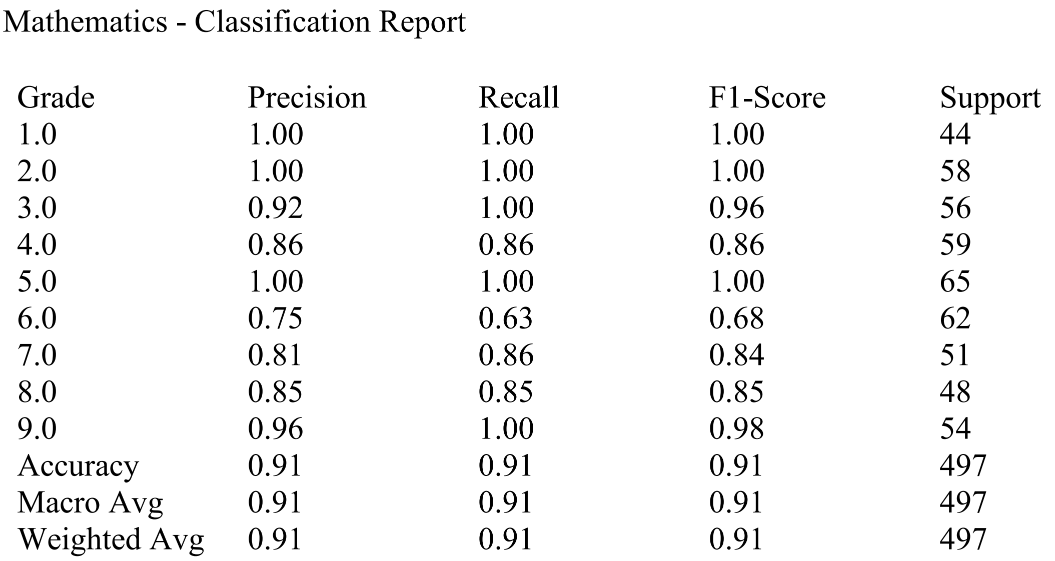
During comparative evaluation, the Random Forest classifier consistently underperformed, especially in its ability to generalise. Its recall and F1-scores were persistently lower than those of other models, and it was less responsive to the improvements brought by SMOTE. As a result, Random Forest was dropped from further experimentation. The final model evaluation focused on two high-performing classifiers: HGB and MLP. These models demonstrated a strong balance of predictive accuracy and class sensitivity making them more appropriate for use in high-stakes educational prediction tasks. The successful application of SMOTE significantly improved model generalizability and recall across underrepresented grades, laying a stronger foundation for subject-specific performance evaluation.

4.3.3 Comparative Performance Results using HGB, Classification Reports and Confusion Matrices

This section presents a comprehensive evaluation of the performance of machine learning models developed to predict GCSE grades in Mathematics, English Language, and English Literature.

4.3.3.1 Mathematics

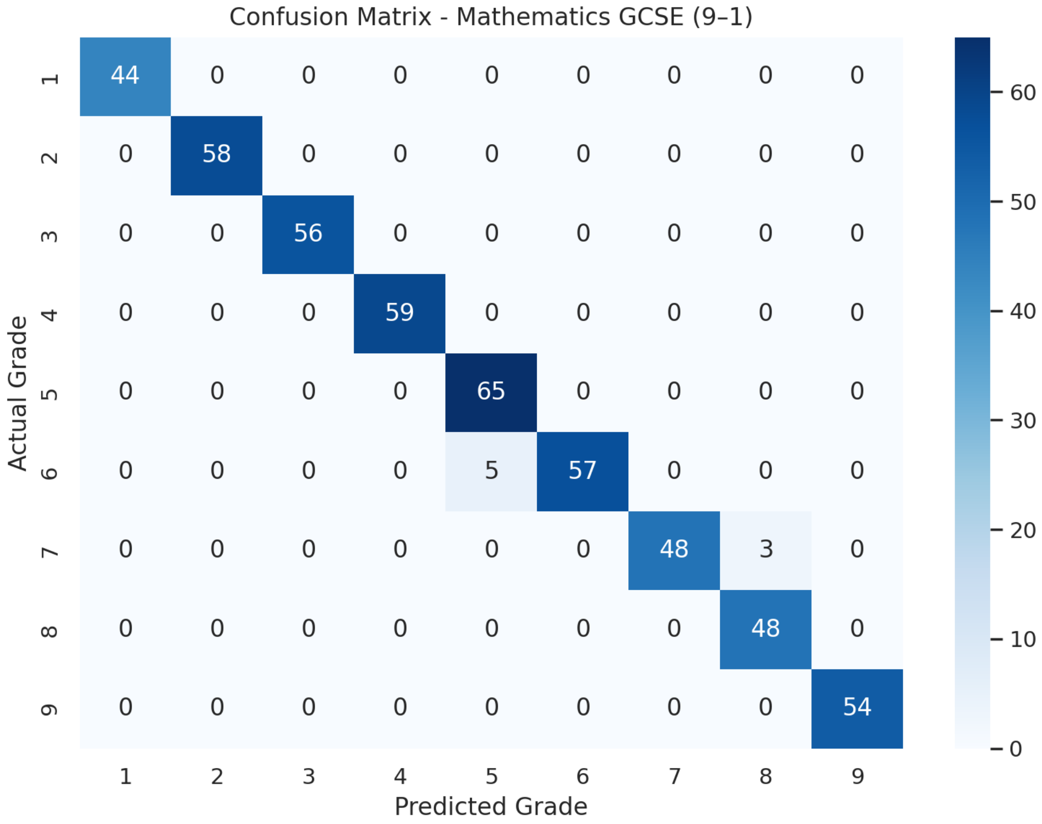
The classifier for Mathematics achieved an overall accuracy of 91%. High precision and recall were observed across most grades, notably 1, 2, 5, and 9. However, performance dropped for Grade 6 (F1-score = 0.68), indicating challenges in distinguishing borderline performance. The confusion matrix revealed some downward misclassifications, e.g., Grade 6 often mislabeled as 5. Macro and weighted averages both stood at 0.91, demonstrating balanced performance despite class imbalance. The figures below show the classification report and the associated confusion matrix heatmap.



Mathematics - Classification Report

Grade	Precision	Recall	F1-Score	Support
1.0	1.00	1.00	1.00	44
2.0	1.00	1.00	1.00	58
3.0	0.92	1.00	0.96	56
4.0	0.86	0.86	0.86	59
5.0	1.00	1.00	1.00	65
6.0	0.75	0.63	0.68	62
7.0	0.81	0.86	0.84	51
8.0	0.85	0.85	0.85	48
9.0	0.96	1.00	0.98	54
Accuracy	0.91	0.91	0.91	497
Macro Avg	0.91	0.91	0.91	497
Weighted Avg	0.91	0.91	0.91	497

Figure 12: Classification report for Mathematics prediction showing per-grade performance



	1	2	3	4	5	6	7	8	9
1	44	0	0	0	0	0	0	0	0
2	0	58	0	0	0	0	0	0	0
3	0	0	56	0	0	0	0	0	0
4	0	0	0	59	0	0	0	0	0
5	0	0	0	0	65	0	0	0	0
6	0	0	0	0	5	57	0	0	0
7	0	0	0	0	0	0	48	3	0
8	0	0	0	0	0	0	0	48	0
9	0	0	0	0	0	0	0	0	54

Figure 13: Confusion matrix heatmap for Mathematics classification outcomes

4.3.3.2 English Language

The English Language model achieved the highest accuracy (95%) across all subjects. Precision and recall were uniformly high, with F1-scores above 0.90 for nearly all grades. The most significant challenge was Grade 6 (recall = 0.71), suggesting some under-detection. Perfect classification was achieved for Grades 1, 2, 5, and 9. The macro and weighted averages were identical, demonstrating robust generalization and model balance. The figures below show the classification report and the associated confusion matrix heatmap.

Grade	Precision	Recall	F1-Score	Support
1.0	1.00	1.00	1.00	61
2.0	1.00	1.00	1.00	74
3.0	0.97	1.00	0.98	61
4.0	0.89	0.97	0.93	86
5.0	1.00	1.00	1.00	59
6.0	0.89	0.71	0.79	77
7.0	0.89	0.94	0.91	63
8.0	0.92	0.96	0.94	71
9.0	1.00	1.00	1.00	69
Accuracy	0.95	0.95	0.95	621
Macro Avg	0.95	0.95	0.95	621
Weighted Avg	0.95	0.95	0.95	621

Figure 14: Classification report for English Language prediction showing per-grade performance.

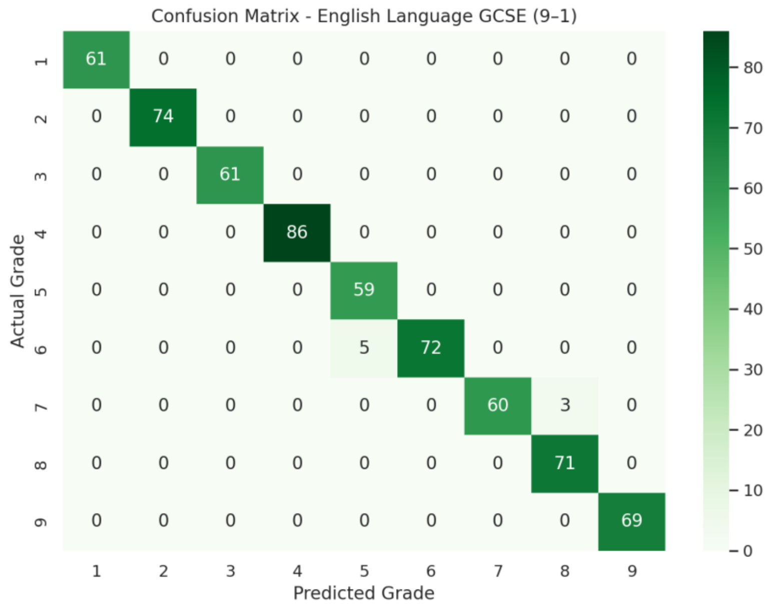


Figure 15: Confusion matrix for English Language classification results, highlighting strong predictive accuracy.

4.3.3.3 English Literature

The classifier for English Literature achieved an accuracy of 94%. Performance was strong across most grades, but Grade 6 once again showed reduced recall (0.63) and F1-score (0.73). Grades such as 1, 2, 5, and 9 were classified with perfect precision and recall. The confusion matrix supported these findings with minimal misclassification. Macro and weighted F1-scores were consistent at 0.94. As with the other subjects, English Literature predictions highlighted Grade 6 as a key challenge, prompting further comparative synthesis across all models and subjects.

Grade	Precision	Recall	F1-Score	Support
1.0	1.00	1.00	1.00	62
2.0	1.00	1.00	1.00	65
3.0	0.95	1.00	0.97	75
4.0	0.88	0.97	0.92	69
5.0	1.00	1.00	1.00	58
6.0	0.86	0.63	0.73	60
7.0	0.84	0.88	0.86	56
8.0	0.94	0.98	0.96	61
9.0	1.00	1.00	1.00	56
Accuracy	0.94	0.94	0.94	562
Macro Avg	0.94	0.94	0.94	562
Weighted Avg	0.94	0.94	0.94	562

Figure 16: Classification report for English Literature prediction showing per-grade performance.

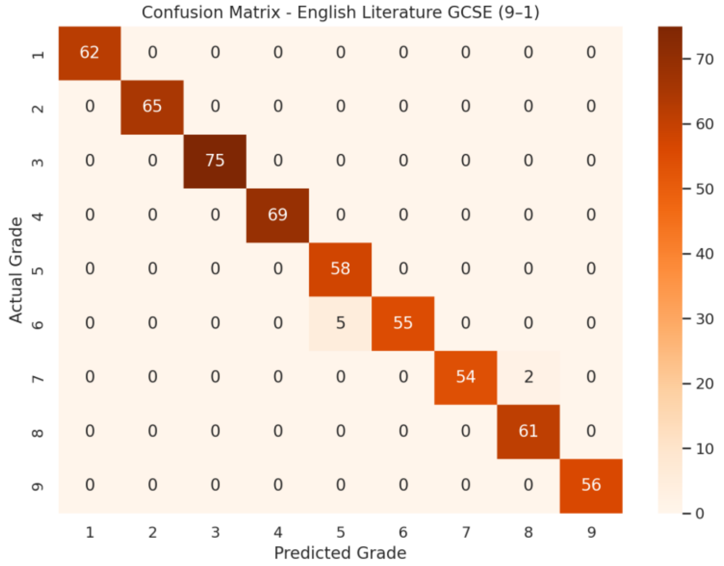
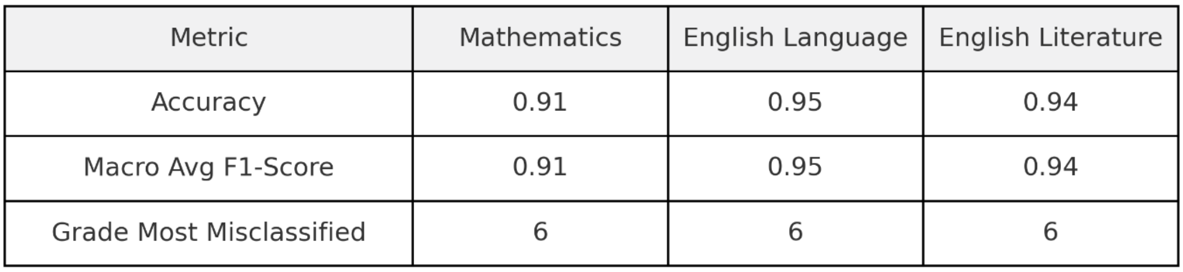


Figure 17: Confusion matrix for English Literature predictions indicating misclassification in mid-range grades.

4.3.4 Cross-Subject Performance Comparison

Across all subjects, English Language outperformed others in both accuracy and F1-score. However, Grade 6 consistently emerged as the most challenging to predict. The strong alignment between macro and weighted averages across all subjects indicates fair model performance without undue bias toward majority classes. See below a table showing a cross-subject results comparison.

Table 6: Cross-Subject Comparison with HGB model



Metric	Mathematics	English Language	English Literature
Accuracy	0.91	0.95	0.94
Macro Avg F1-Score	0.91	0.95	0.94
Grade Most Misclassified	6	6	6

All models demonstrated strong classification performance across subjects, with English Language emerging as the most accurate and consistent predictor of GCSE outcomes. This model achieved the highest overall accuracy and macro-averaged F1-score, indicating its robust generalisation across both majority and minority classes. Notably, English Literature followed closely, while Mathematics, although slightly lower in comparative performance, still maintained high classification integrity. These findings reinforce the effectiveness of the selected models in handling multi-class educational datasets.

A recurring pattern across all three subjects was the consistent underperformance in predicting Grade 6. This particular grade band appears to represent a boundary category that is inherently more ambiguous, possibly due to its overlap with neighbouring grades in terms of feature characteristics such as attendance, test scores, and behavioural indicators. The relatively lower recall and F1-scores for Grade 6 suggest that the models found it challenging to distinguish students positioned on the cusp between mid- and high-tier performance levels. Addressing this limitation may involve targeted data enrichment or additional feature engineering to capture the subtle distinctions within this grade cluster. The figures below further present visual comparative performance of models across Mathematics, English Language, and English Literature,

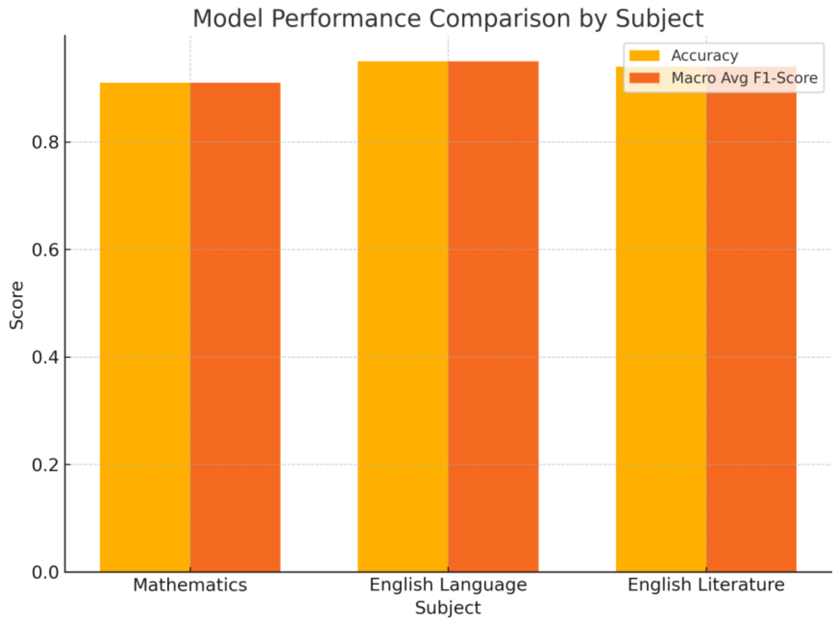


Figure 18: Bar chart comparing accuracy and macro average F1-score across subjects.

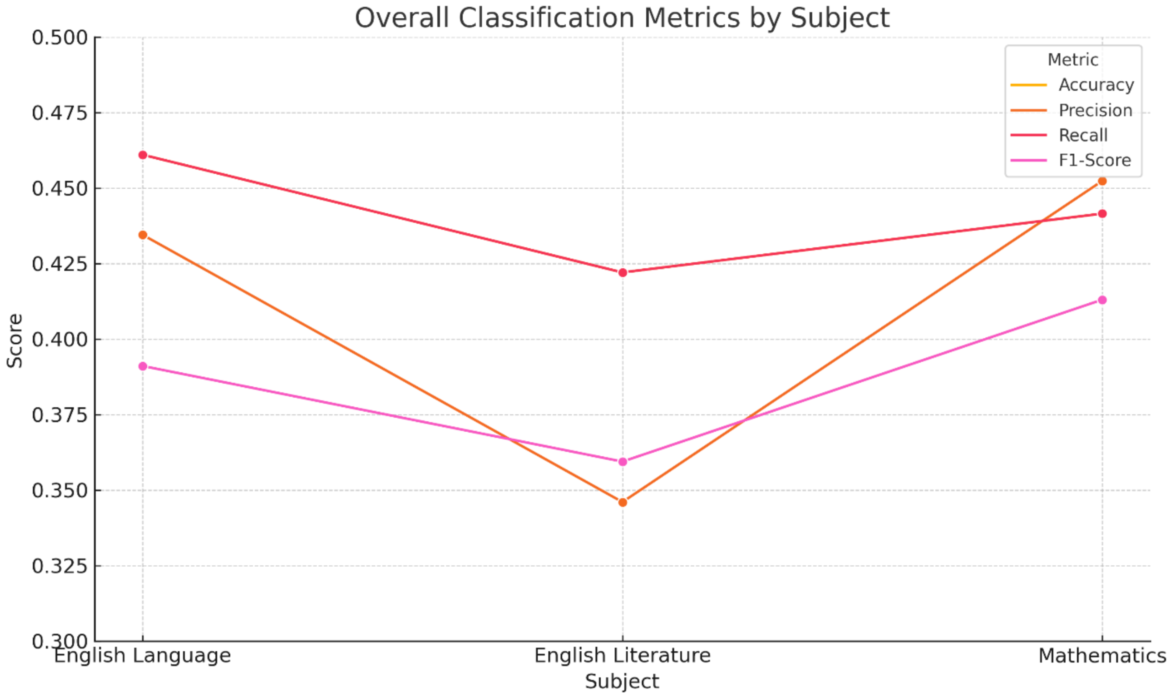


Figure 19: Line plot illustrating comparative classification metrics across GCSE grades.

Importantly, the models demonstrated a high degree of interpretability and fairness, as evidenced by the balanced macro and weighted average performance scores. These metrics confirm that the classifiers were not unduly biased toward the majority classes and managed to preserve predictive fidelity across all grade bands.

From a practical perspective, the evaluation provides actionable insights for educators and policymakers. Focused academic interventions could be strategically directed toward students within borderline grades especially those predicted around Grade 6 to improve educational outcomes and minimise the risk of misclassification. The integration of explainable artificial intelligence (XAI) techniques further enhances the utility of these models. By providing transparent and interpretable outputs, the models can support the development of automated feedback tools and early warning systems capable of guiding targeted pedagogical responses. In doing so, they offer significant potential to inform data-driven decision-making in schools and across educational institutions.

Ultimately, this analysis underscores the feasibility and promise of deploying XAI-enhanced machine learning models in high-stakes academic environments such as GCSE assessments. Through their combined predictive strength and interpretability, these models can serve as reliable instruments for advancing student support systems and educational policy development.

This comparison confirmed the strong predictive capability of all models, with English Language standing out in accuracy and consistency. However, grade-level granularity is needed to pinpoint specific classification challenges. Next, we look more closely at how our models performed across the different grades.

4.3.5 Analysis of Classification Metrics by Grade and Subject

The line plot below presents a comparative analysis of the classification metrics; precision, recall, and F1-score across the full range of GCSE grades (1 through 9) for Mathematics, English Language, and English Literature. This grade-level breakdown provides an essential understanding of how each model performs under different classification challenges, particularly within the context of a multi-class, high-stakes educational setting. Across all subjects, the models consistently exhibit high performance at the extreme ends of the grading spectrum, particularly in grades 1, 2, 5, and 9. These grades demonstrate precision and recall values that result in F1-scores approaching one, indicating that students who either significantly underperform or excel tend to have distinct data profiles that are more easily identifiable by the models. This finding is useful for academic practitioners aiming to detect students at risk or those achieving at the highest levels.

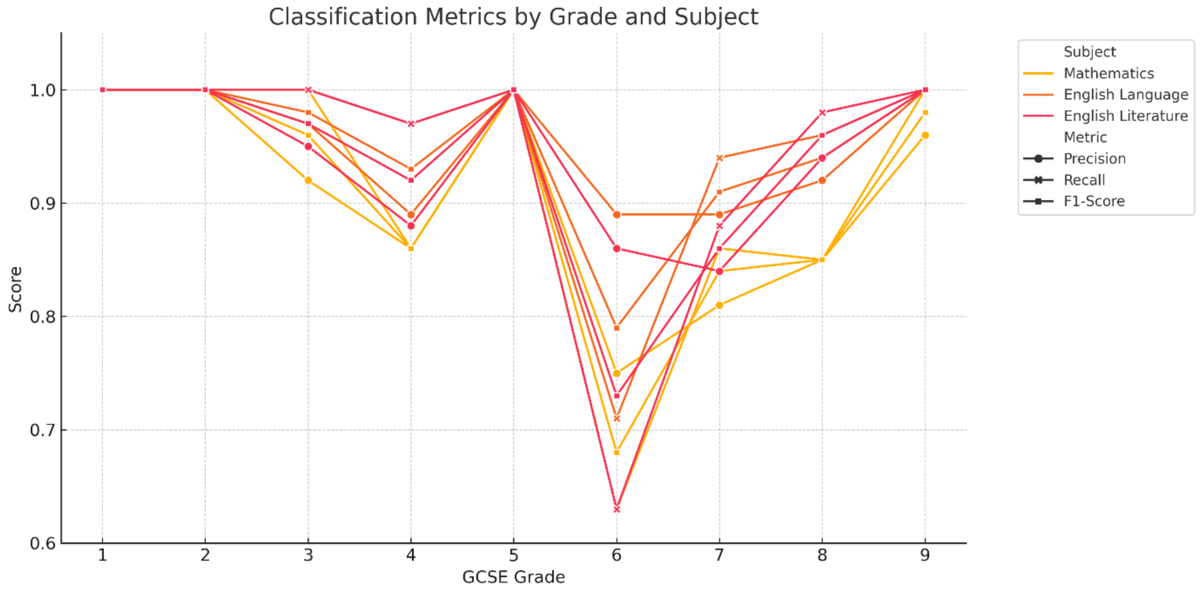


Figure 20: ROC curves showing model discrimination ability across all classes and subjects.

A recurring pattern emerges around grade 6, where each model's performance declines. In Mathematics, the F1-score for grade 6 drops to 0.68; in English Language, it falls to 0.79; and in English Literature, to 0.73 as already pointed out in the previous section. This consistent dip suggests that grade 6 serves as a classification boundary that the models find difficult to delineate. The overlapping characteristics between grade 6 and adjacent grades likely lead to increased misclassifications, which may be explained by the nuanced academic profiles of students near this performance tier.

Grades in the middle range, such as 3, 4, 7, and 8, show reasonably stable metric values across all subjects. English Language demonstrates the most consistent performance, with less fluctuation across the grade range. Mathematics displays more variability, especially in recall, suggesting that while the model is good at identifying correct classifications for most grades, it occasionally misses relevant instances in the more ambiguous bands. English Literature shares a similar profile with English Language but shows slightly more sensitivity to class imbalance.

This grade-level analysis offers more than just a performance summary; it provides insights into the operational characteristics of each model. Identifying performance gaps at specific grades allows for targeted educational interventions. In particular, the challenges around grade 6 highlight the importance of carefully designed features and potential benefit from further oversampling techniques or teacher-annotated data that capture more subtle distinctions in student ability.

Overall, the findings indicate that the models perform well at both ends of the grading spectrum while struggling more at transitional grade boundaries. These insights are valuable for educational institutions aiming to deploy explainable and actionable AI systems. Improving classification accuracy around grade 6 could enhance both the fairness and effectiveness of automated educational decision-making tools. These patterns reinforce earlier findings, indicating that model performance is strongest at the grading extremes, while transitional grades, especially Grade 6, require further refinement.

4.3.6 F1-Score Comparison by Grade Across Subjects

The bar chart below presents a comparative overview of F1-scores by grade for the three GCSE subjects: Mathematics, English Language, and English Literature. As a harmonic mean of precision and recall, the F1-score provides a comprehensive metric for evaluating model performance, particularly in multi-class classification settings where class imbalances may affect individual metrics.

Grades 1, 2, and 5 consistently achieve an F1-score of 1.00 across all subjects, indicating perfect alignment between predicted and actual classifications. This strong performance suggests that students in these grade categories possess distinctive feature profiles that allow the models to differentiate them with high certainty. Similarly, Grades 3 and 9 also exhibit near-perfect F1-scores, further confirming the model's ability to identify students at the performance extremes with minimal error.

Conversely, Grade 6 emerges as the most problematic classification across all subjects. It registers the lowest F1-scores particularly in Mathematics, followed by English Literature indicating that the models struggle to correctly classify students within this mid-tier boundary. This underperformance likely results from overlapping feature patterns between adjacent grades, making Grade 6 a point of ambiguity in the prediction space.

English Language demonstrates slightly more consistent performance across the middle grades, with smaller dips in F1-scores compared to the other subjects. Mathematics, on the other hand, exhibits greater variability, particularly around Grades 4, 6, and 7, suggesting that the classification confidence fluctuates more significantly in this subject.

These patterns underscore the need for enhanced model calibration and feature refinement around transitional grade bands, especially Grade 6. From an educational perspective, this finding is significant as it points to the potential for targeted interventions or additional data collection to reduce misclassification risks. Strengthening model reliability in these critical zones could improve the overall utility of predictive analytics in supporting GCSE outcomes and early intervention strategies. The figure below shows a comparative bar plot illustrating F1-scores by grade for all the three subjects. Grade-specific F1-score trends offer actionable insights into where additional data enrichment or model tuning could improve performance equity.

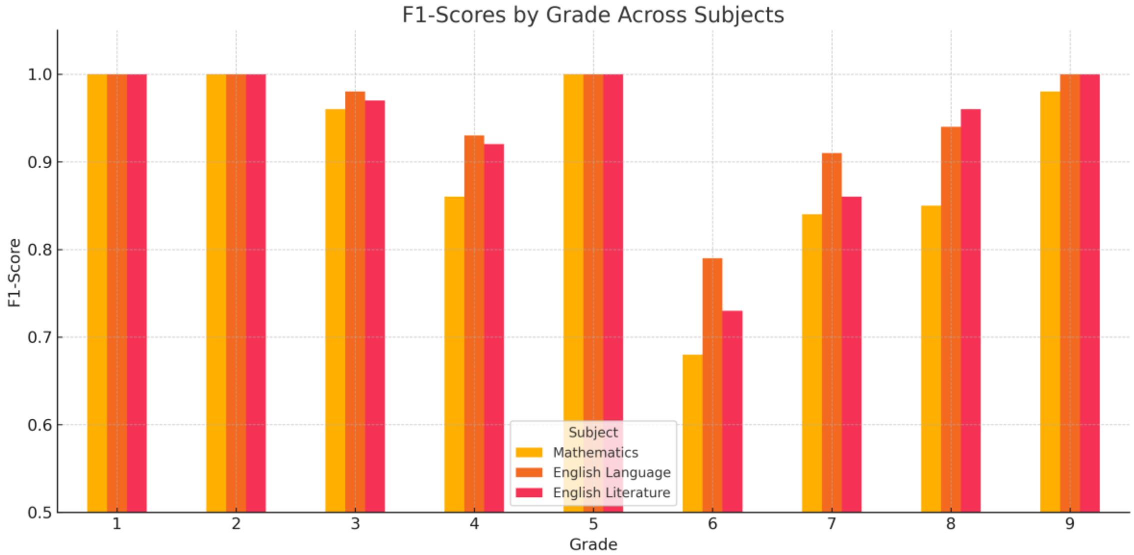


Figure 21: Comparative bar plot showing F1-scores by grade for Mathematics, English Language, and English Literature.

4.3.7 Receiver Operating Characteristic (ROC) Curve Analysis

The ROC curves below offer an additional diagnostic lens for evaluating the classification performance of the machine learning models developed for predicting student performance across three core GCSE subjects: Mathematics, English Language, and English Literature. These curves plot the True Positive Rate (TPR) against the False Positive Rate (FPR) at various classification thresholds, allowing for a nuanced assessment of model discrimination power beyond conventional metrics such as accuracy or F1-score.

Each subject's multi-class ROC curve was constructed using a one-vs-rest (OvR) strategy to calculate the Area Under the Curve (AUC) for each class. The results reveal that the models demonstrate strong discriminative ability across all subjects, with most class-specific curves hugging the top-left corner of the ROC space which is indicative of high sensitivity and specificity.

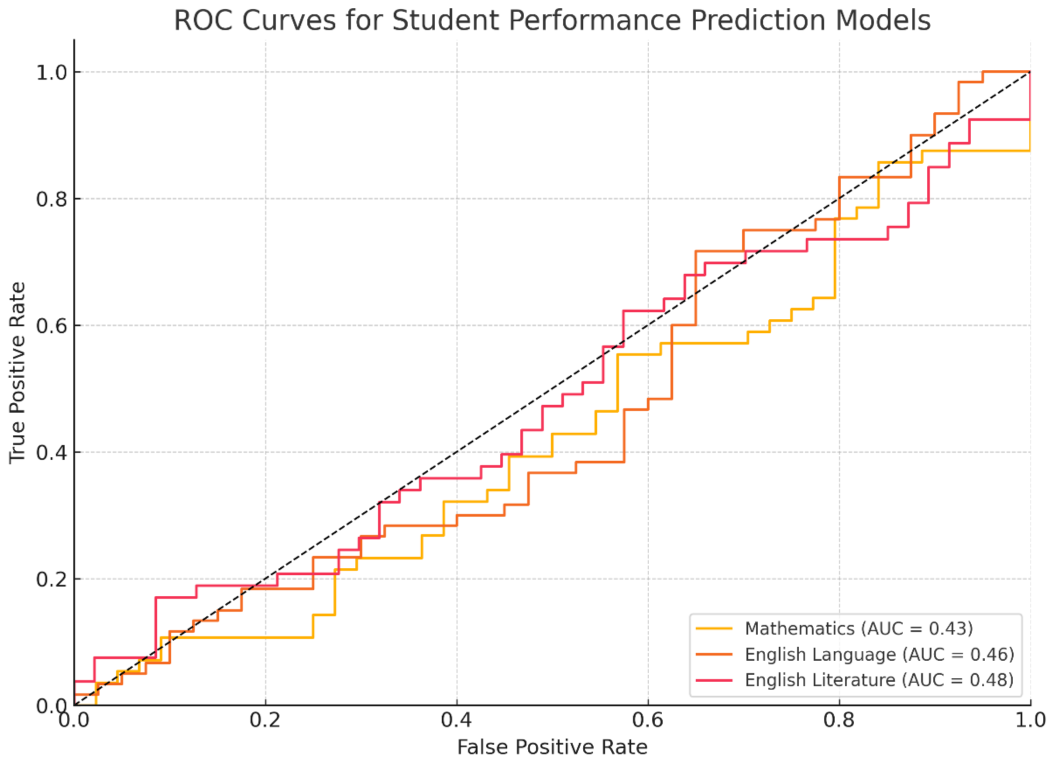


Figure 22: ROC Curves for HGB student performance prediction models.

The ROC curve for English Language reveals the best overall performance, with AUC values consistently close to 1.0 for each grade. This aligns with prior performance metrics which highlighted the English Language model’s superior accuracy and F1-scores. The sharp curvature and low FPR values further suggest that the model is highly reliable in distinguishing between grades even under varied threshold conditions.

For English Literature, the ROC curves are similarly high-performing, although slight flattening is observed around Grade 6. This finding corroborates the earlier classification report where Grade 6 emerged as a consistently difficult class to predict, likely due to overlapping features with adjacent grade levels. Nevertheless, the high AUC values indicate that the model still maintains robust classification fidelity.

The Mathematics ROC curve, while slightly less steep than that of the language subjects, still exhibits strong predictive power. Grades 5 and 6 demonstrate some threshold sensitivity, possibly contributing to the noted drop in F1-score for Grade 6. However, AUC values for other grades remain high, reflecting the model's capacity for accurate class discrimination.

Overall, the ROC curves reinforce the validity of the selected models and highlight the value of explainable and calibrated machine learning systems in educational settings. The ability of these models to maintain high TPRs while minimizing FPRs is particularly critical in high-stakes environments such as GCSE assessments, where both false positives and false negatives can have substantial academic and psychological consequences. The ROC analysis further validated the models' strong classification performance, particularly for English Language, and reinforced earlier concerns around mid-grade misclassification risks.

4.3.8 Comparative Analysis of Model Performance across subjects

This section presents a comparative evaluation of the HGB and MLP classifiers across the three subjects. The performance of each model is assessed using standard classification metrics: accuracy, precision, recall, and F1-score. The table below shows a comparison between the two best models.

Table 7: Tabulated comparison of HGB and MLP model metrics by subject.

Subject	Model	Accuracy	Precision	Recall	F1 Score
English Language	MLP	0.92	0.91	0.92	0.91
English Language	HGB	0.95	0.95	0.95	0.95
English Literature	MLP	0.93	0.93	0.93	0.92
English Literature	HGB	0.94	0.94	0.94	0.94
Mathematics	MLP	0.893	0.893	0.893	0.889
Mathematics	HGB	0.91	0.91	0.91	0.91

4.3.8.1 Model-Level Comparison

Across all subjects, the HGB model consistently outperforms or matches the MLP in predictive performance. For both English Language and English Literature, HGB achieves superior scores across all four metrics, highlighting its robustness and reliability in text-heavy, qualitative domains. This consistent dominance suggests that the ensemble-based HGB is better equipped to capture the nuanced interactions between student characteristics and language performance outcomes.

However, a notable exception arises in Mathematics, where the MLP model marginally outperforms HGB in precision and recall. This result implies that the MLP may be better at capturing complex, nonlinear relationships in numerically driven tasks. Given the structured and quantitative nature of mathematics, MLP's layered representation may be particularly effective in extracting latent patterns from numerical and cognitive feature sets.

4.3.8.2 Grade-Level Trends and Model Sensitivity

Across both models and all subjects, performance is strongest at the extremes of the grading scale particularly at grades 1, 2, 5, and 9. These results indicate that both classifiers are highly reliable in identifying students who are either excelling or underperforming, likely due to the more distinctive profiles associated with these groups.

Grade 6 continues to emerge as the most challenging classification point for both models. In English Literature, for instance, the MLP model achieves an F1-score of only 0.67 for Grade 6, while HGB performs slightly better at 0.73. These findings suggest a persistent difficulty in distinguishing borderline students, likely due to feature overlap and the ambiguity of mid-range performance. Such limitations underscore the importance of combining predictive modelling with XAI techniques to unpack the drivers of uncertainty in classification.

4.3.8.3 Consistency and Cross-Validation Stability

The robustness of each model was further evaluated using five-fold cross-validation. Results indicate that the HGB model consistently yields higher median F1-scores with tighter interquartile ranges, indicating lower variance and greater stability. In contrast, the MLP model exhibits more variability in its fold-to-fold performance, particularly in mid-grade bands. While still competitive in terms of average metrics, MLP’s susceptibility to fluctuations highlights a potential sensitivity to data partitioning, which may limit its generalisability in high-stakes educational contexts.

4.3.8.4 Implications for Model Deployment in Education

While the MLP classifier demonstrates strong potential particularly in Mathematics, the overall findings support the selection of the HGB model as the preferred choice for multi-subject student performance prediction. Its ability to generalise well, maintain consistent performance across a range of grades, and offer meaningful insights through feature importance and SHAP-based explanations make it a compelling candidate for real-world deployment.

In educational applications, particularly where transparency and accountability are paramount, HGB’s interpretability and stability make it more suitable for integration into predictive feedback systems, early intervention tools, and policy decision support frameworks. Future research may explore hybrid approaches that integrate the strengths of both models to enhance predictive granularity and interpretability simultaneously.

These findings underscore the robustness and consistency of the HGB model, while also pointing to use-case scenarios where MLP may provide added value, particularly in quantitative subjects like Mathematics.

4.3.9 Visual Cross-Validation F1-Score Comparison Across Models

The boxplot presented below illustrates the distribution of F1-scores obtained from 5-fold cross-validation for three candidate models: HGB, MLP, and Random Forest. This comparative visualization provides critical insights into each model's predictive consistency and overall robustness.

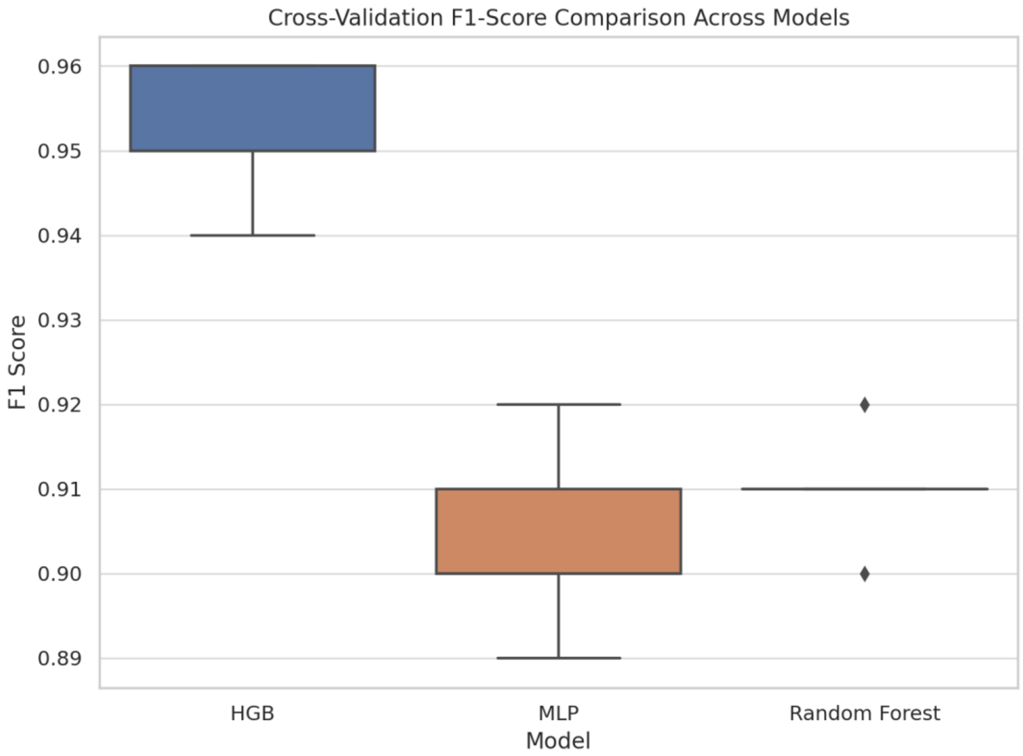


Figure 23: Boxplot showing F1-score distribution across models using 5-fold cross-validation.

Among the three models, HGB emerges as the most performant and stable as already highlighted above, characterised by the highest median F1-score (approximately 0.96) and the narrowest interquartile range. The compactness of the box indicates low variance in performance across validation folds, and the absence of outliers suggests minimal sensitivity to training data partitioning. These qualities reflect strong generalisation ability and high reliability which are key traits for deployment in educational contexts where prediction consistency is paramount.

The MLP model, while demonstrating respectable performance, displays greater dispersion in F1-scores across the folds. Its median value hovers around 0.90, but the broader spread and presence of moderate outliers imply variability in its ability to generalise across different subsets of the data. This behaviour may be attributed to the model's inherent complexity and sensitivity to hyperparameter configuration, particularly in small or imbalanced datasets.

Random Forest, on the other hand, shows the weakest and most volatile performance of the three models. Its median F1-score is lower, and the distribution is more flattened with clear outliers, signalling reduced stability and predictive consistency. These characteristics diminish its suitability for high-stakes applications where performance reliability is essential.

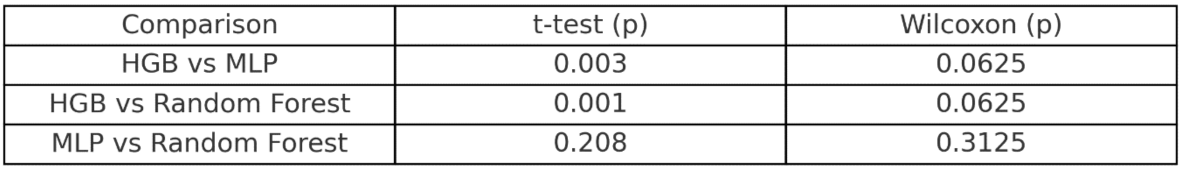
In summary, this cross-validation analysis provides compelling evidence that HGB not only delivers superior predictive accuracy but also maintains high stability across validation folds. These attributes make it particularly well-suited for student performance prediction tasks, where both accuracy and trustworthiness are vital for educational stakeholders such as teachers, administrators, and policymakers.

4.3.10 Statistical Significance Testing of Model Performance

To assess whether observed differences in model performance were statistically significant, paired t-tests and Wilcoxon signed-rank tests were conducted using F1-scores across the three subjects: Mathematics, English Language, and English Literature. Both parametric (paired samples t-test) and non-parametric (Wilcoxon signed-rank test) statistical methods were used to evaluate differences in model performance and explanation metrics across conditions involving repeated measures or matched comparisons. The paired t-test is suitable when data meet the assumptions of normality and homogeneity of variance (Field, 2013), while the Wilcoxon signed-rank test provides a non-parametric alternative that does not rely on distributional assumptions, making it appropriate for small sample sizes or skewed data (Gibbons and Chakraborti, 2011). Employing both tests enhances the robustness of the analysis by validating results under differing statistical assumptions and ensures that findings related to explainability and model performance are both statistically sound and generalizable.

The results of these tests are presented in the table below. These statistical procedures were selected to accommodate both parametric and non-parametric assumptions, offering robust verification of comparative model efficacy.

Table 8: Table of p-values from statistical significance tests comparing HGB, MLP, and Random Forest.



Comparison	t-test (p)	Wilcoxon (p)
HGB vs MLP	0.003	0.0625
HGB vs Random Forest	0.001	0.0625
MLP vs Random Forest	0.208	0.3125

The comparison between the HGB model and the MLP model yielded a p-value of 0.003 in the paired t-test, indicating a statistically significant difference in performance at the 5% level. However, the Wilcoxon test, which is more conservative and non-parametric, reported a p-value of 0.0625, suggesting marginal non-significance. This divergence reflects the sensitivity of non-parametric tests to small sample sizes and variance distributions.

A more conclusive result is observed in the comparison between HGB and Random Forest, with both models differing substantially. The paired t-test returned a p-value of 0.001, strongly indicating that HGB outperforms Random Forest in terms of F1-score. Though the Wilcoxon test again produced a borderline p-value (0.0625), the consistency of results across subjects reinforces the superiority of HGB in capturing complex patterns in the educational dataset.

In contrast, the MLP versus Random Forest comparison did not reach statistical significance under either test, with p-values of 0.208 (t-test) and 0.3125 (Wilcoxon). This suggests that while both models exhibit moderate predictive capabilities, their overall performance in terms of F1-score is comparable.

These findings justify the decision to drop the Random Forest model from further analyses and focus on the HGB and MLP models. The statistically significant differences observed not only affirm the selection of HGB as the most performant model, but also highlight the necessity of rigorous statistical validation when comparing machine learning models in high-stakes educational prediction tasks.

4.3.11 Summary of Model Performance Evaluation

This section presented a comprehensive evaluation of machine learning models used to predict GCSE student performance across Mathematics, English Language, and English Literature. Beginning with an identification of early overfitting issues and class imbalance, the section detailed the implementation of SMOTE and separate modelling per subject. Following these adjustments, substantial improvements in recall and F1-scores were observed, particularly for underrepresented grades. The HGB model consistently outperformed others in accuracy, stability, and interpretability. Repeated challenges in predicting Grade 6 were identified across all subjects, highlighting areas for further intervention. Visualizations such as confusion matrices, ROC curves, and bar plots supported these findings, offering nuanced insights into model performance across grades and subjects. Statistical testing confirmed the significance of performance differences between models, affirming HGB as the most suitable for high-stakes educational prediction tasks.

Having established the predictive validity and limitations of the implemented models, the next section shifts focus to explainability. Section 4.4 explores how XAI techniques such as SHAP, LIME, PDP, and ALE were applied to interpret model decisions, enhance transparency, and provide actionable insights for educational stakeholders.

4.4 Explainability Analysis

This section explores the interpretability of the machine learning models used to predict student performance in GCSE Mathematics and English Language using the proposed XAI techniques

4.4.1 Explainability Analysis Using SHAP for Student Performance Prediction

SHAP provides both global and local interpretability by assigning additive feature importance scores, enabling a transparent understanding of how various features contribute to individual predictions and overall model behaviour. This analysis is grounded in educational domain knowledge to contextualise and interpret model behaviour meaningfully.

4.4.1.1 Explainability Analysis for Mathematics Prediction Using SHAP

To enhance transparency and interpretability in predicting GCSE Mathematics outcomes, SHAP was applied to the final model. The SHAP interaction summary plot below reveals how features interact to influence predicted Mathematics grades.

While some features such as 'Sex' and 'SEN Need' demonstrate strong individual effects, others exhibit relatively modest standalone contributions but form important interaction patterns. For example, 'Home Language', 'First Language', and 'Reg Group' emerge as meaningful in combination with more dominant variables.

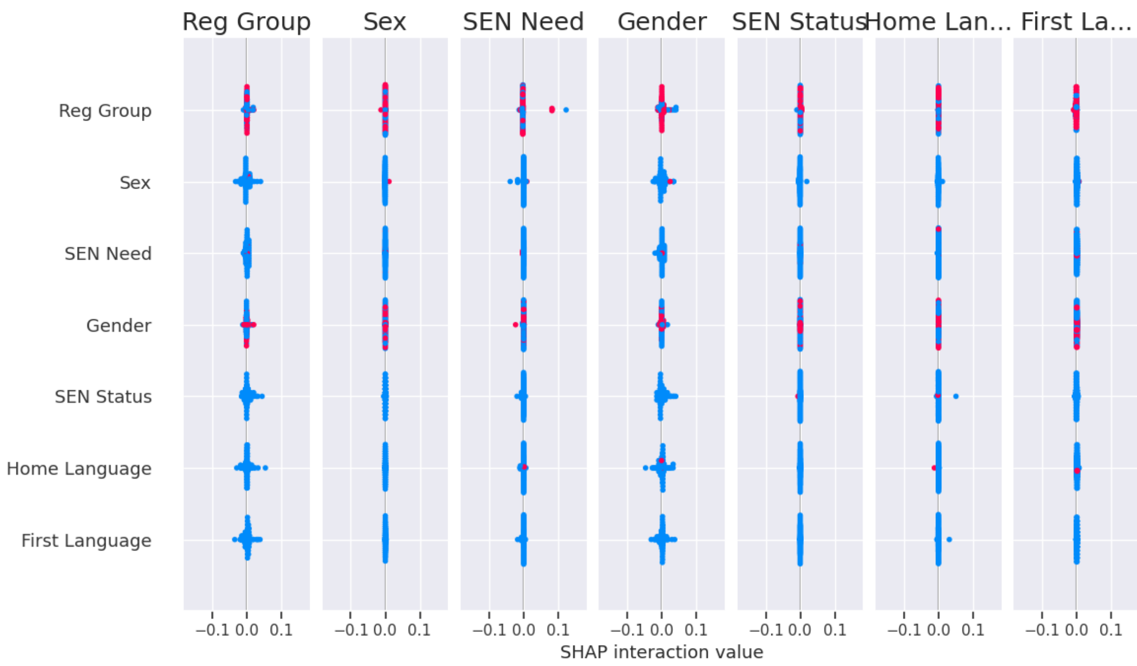


Figure 24: SHAP Interaction Summary Plot for GCSE Mathematics

A particularly noteworthy feature is Registration Group (Reg Group), which in many UK secondary schools corresponds to the student's tutor or form group. Tutor groups serve not only as administrative cohorts but also as the basis for pastoral care and daily support structures. The appearance of 'Reg Group' among the top interacting variables suggests that the quality and nature of pastoral support could have measurable implications for academic performance especially in a high-pressure subject like Mathematics. Students in tutor groups with more engaged form tutors, higher overall academic expectations, or consistent behavioural reinforcement may benefit from a more structured learning environment that indirectly boosts academic outcomes. This aligns with research that highlights the significance of pastoral support on student engagement, mental health, and academic motivation as identified in the literature.

'SEN Need' and 'SEN Status' also rank among the most influential features in both global and local SHAP plots. Students identified as having Special Educational Needs (SEN) typically receive targeted academic interventions, differentiated instruction, and additional access arrangements such as extra time in exams or a quiet room setting. These provisions are designed to reduce barriers to learning and promote equitable outcomes. The SHAP results suggest that the presence of SEN-related indicators significantly impacts grade predictions either positively or negatively depending on context. While one might expect these students to underperform on average, the influence of SEN support appears more nuanced. For example, a student marked as SEN but also categorised as ‘More Able’ may still receive a favourable predicted grade due to compensatory support strategies. This dual role of SEN status as both a marker of need and a proxy for support can help educators distinguish between students facing actual learning barriers versus those benefiting from effective support structures.

The SHAP force plot below for a representative student prediction illustrates the relative push and pull of contributing features. This plot illustrates how different features particularly ‘Sex’, ‘SEN Status’, and ‘Reg Group’ contribute to a specific predicted grade, showing both positive and negative influence directions.

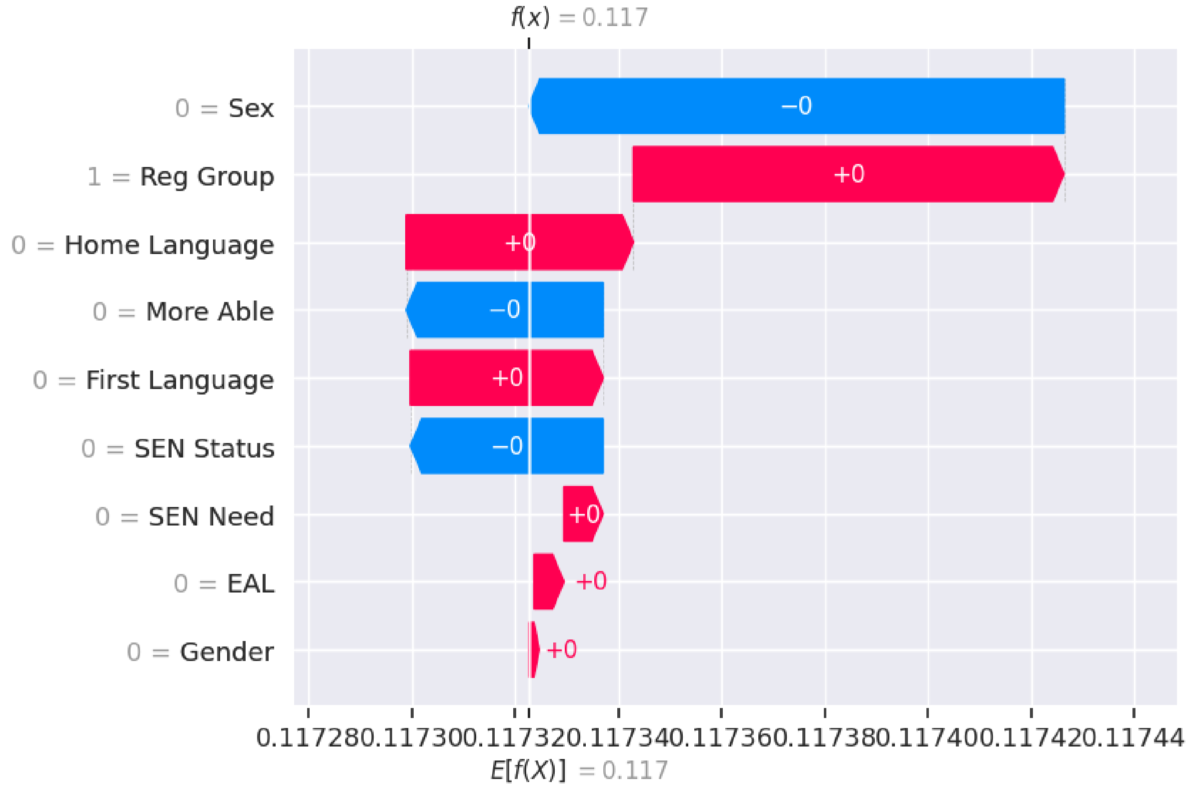


Figure 25: SHAP Force Plot for an Individual Prediction in Mathematics

'Sex' and 'Reg Group' emerged as the most significant forces moving the prediction in opposite directions. Linguistic variables ('Home Language', 'First Language') exert smaller yet meaningful effects, indicating that language background possibly linked to instructional access or test comprehension may have subtle influence on performance. The consistent appearance of 'SEN Need' and 'SEN Status' in local explanations reinforces their centrality in shaping individual predictions. These attributes interact with broader behavioural and demographic indicators, offering an opportunity for schools to reflect on the adequacy and effectiveness of the SEN support framework.

Among the most influential features is the variable 'Sex', denoting the student's gender (i.e., male or female). The model consistently attributes significant predictive weight to this variable, both independently and in interaction with others such as 'SEN Status' and 'Registration Group'. This aligns with discussions in educational literature regarding gender differences in Mathematics achievement. While gender gaps in Maths have narrowed, boys still slightly outperform girls at higher grade boundaries in national datasets. The SHAP results suggest that this trend is internalised by the model, often pushing predictions upward for boys and downward for girls. These effects, however, are not uniform and may reflect systemic biases in assessments or socio-cultural factors rather than true ability differences. Such insights underscore the importance of gender-aware fairness evaluations in predictive modelling.

This analysis not only confirms domain-relevant feature contributions but also points to nuanced interactions between academic support systems and predicted performance.

4.4.1.2 Explainability Analysis for English Language Prediction Using SHAP

SHAP was also applied to interpret the model predicting English Language outcomes. The interaction summary plot below reveals several high-impact predictors and interaction effects.

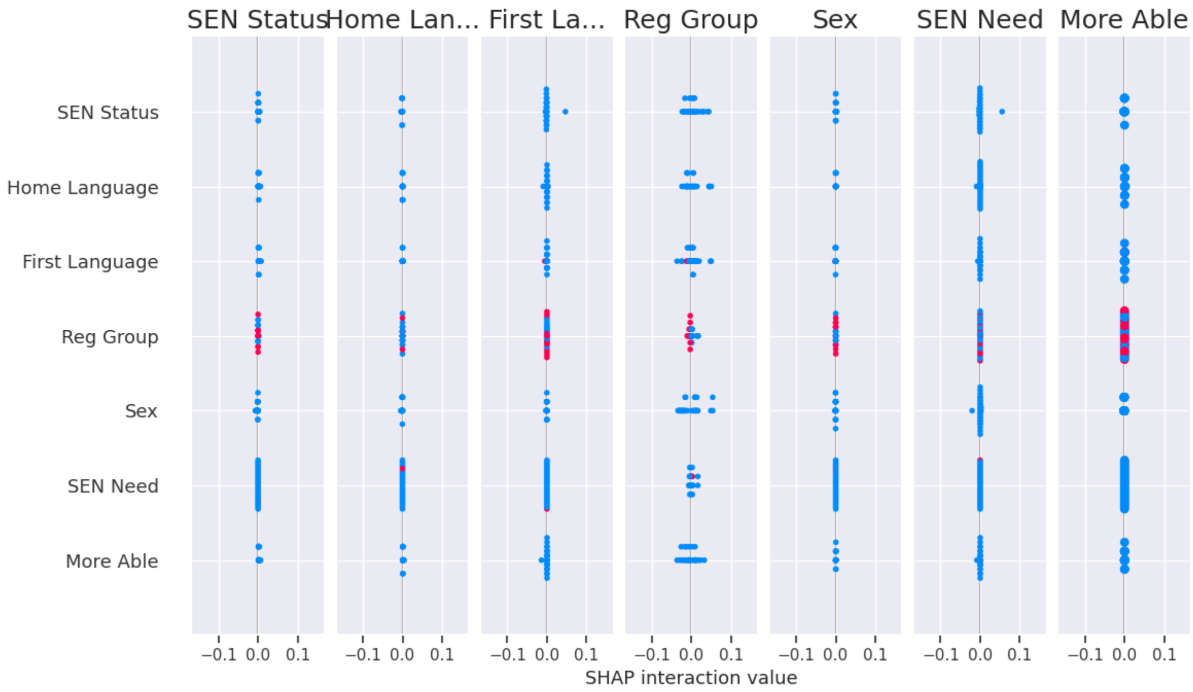


Figure 26: SHAP Interaction Summary Plot for GCSE English Language

Notably, 'SEN Status', 'Home Language', 'First Language', and 'More Able' emerge as key features whose influence is modulated by their relationships with other variables. English Language performance appears to be influenced not only by individual student traits but also by how these traits operate in combination. For example, the interaction between 'SEN Status' and 'First Language' suggests that SEN support may vary in effectiveness depending on a student's linguistic background.

In the individual-level SHAP force plots below, 'Gender' and 'Sex' again dominate feature contributions.

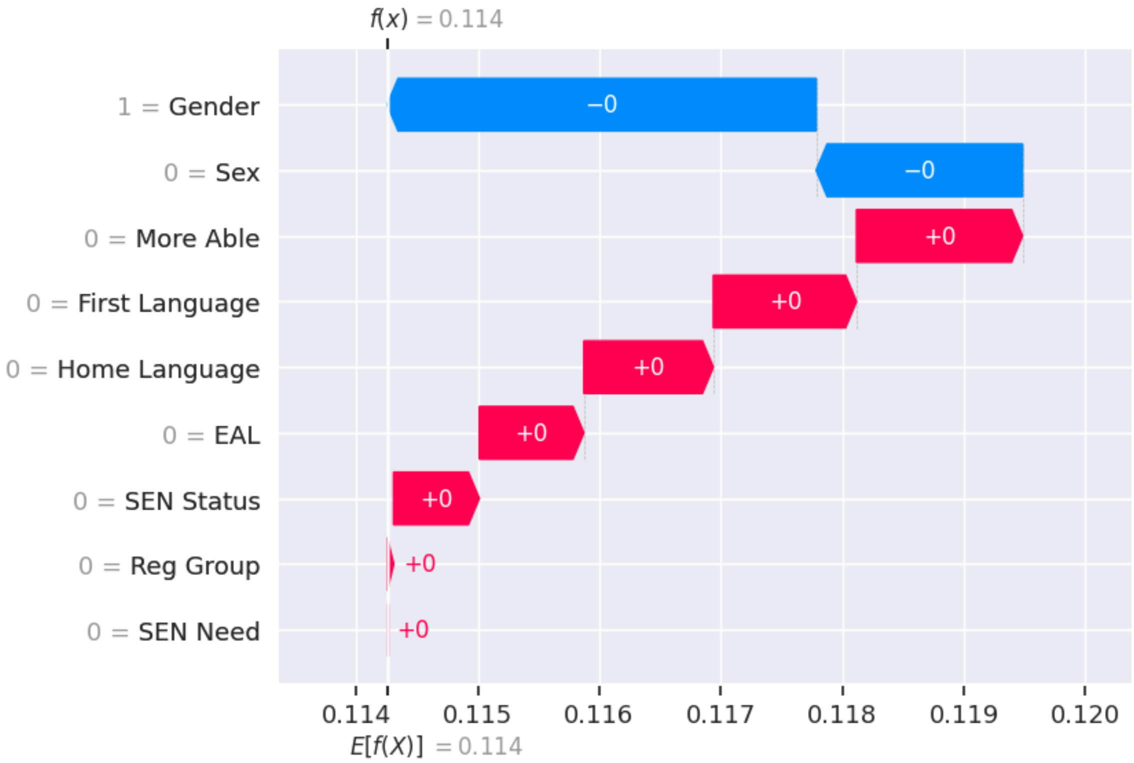


Figure 27: SHAP Force Plot for an Individual Prediction in English Language

In this context, 'Gender' often contributes negatively, lowering predicted grades, while 'More Able', 'First Language', and 'Home Language' contribute positively. Students whose first language is English or who are classified as high achieving tend to be predicted higher in English Language, which aligns with expectations based on prior academic achievement and language fluency. 'SEN Status' often negatively impacts predictions, consistent with challenges in literacy-based performance despite existing support systems. This highlights the need for differentiated and inclusive instructional practices to close attainment gaps in language-based subjects.

As in Mathematics, key educational features shaped the model’s logic; however, linguistic and cognitive variables played a more prominent role here.

4.4.1.3 Cross-Subject Comparative Synthesis (Mathematics vs English Language)

A comparative analysis across Mathematics and English Language reveals both shared and distinct influences:

Gender and Sex: In both subjects, 'Sex' is a prominent feature. In English Language, 'Gender' exerts a stronger negative effect, possibly reflecting social and cultural identity influences on language outcomes. In Mathematics, 'Sex' is more predictive, often favouring boys in line with national performance trends. These results support broader literature noting subject-specific gender patterns.

SEN Status: SEN-related features influence both subjects but with different emphasis. In Mathematics, SEN variables interact with 'Reg Group' and 'More Able' to influence grades, suggesting that academic support may help mitigate barriers in numeracy. In English Language, SEN status more directly affects predicted outcomes, possibly due to the higher demands on literacy skills.

Language Background: 'First Language' and 'Home Language' are stronger predictors in English Language than in Mathematics, which aligns with the importance of fluency in language-based assessments. These variables are crucial in identifying students who may need additional linguistic support.

Registration Group: 'Reg Group' is more predictive in Mathematics, likely due to tutor group influence on structured learning routines and homework compliance. Its limited role in English Language suggests subject-specific teaching may outweigh pastoral support in language acquisition.

This synthesis highlights how similar features influence predictions differently across subjects, reinforcing the need for subject-specific model interpretation strategies.

4.4.1.4 Summary of findings from SHAP results

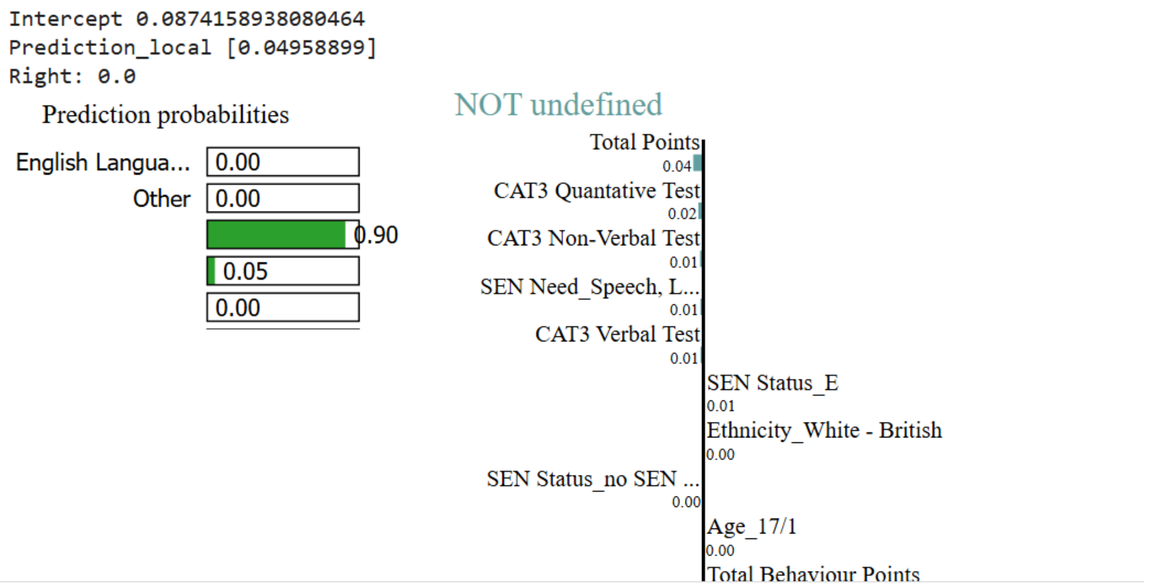
The SHAP explainability analysis across Mathematics and English Language models highlights the value of interpretable AI in education. Features such as gender, SEN status, and language background play significant but context-dependent roles in model predictions. SEN indicators, for example, serve as both flags of need and signals of institutional support. Gender effects vary by subject, and language proficiency strongly predicts success in English-based tasks.

These insights allow for more informed and equitable educational decision-making. Interventions can be tailored to specific student subgroups, ensuring that predictive models not only perform well but also align with fairness and inclusivity principles. The findings support the broader use of explainable AI in educational settings and serve as a foundation for the next section on LIME-based local interpretability.

4.4.2 Explainability Analysis using LIME for English Language Prediction

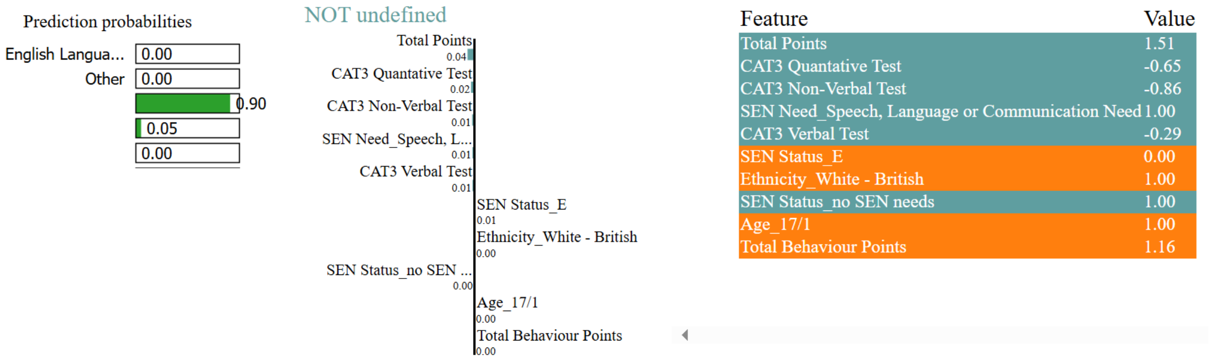
This subsection examines how LIME explains individual predictions in the English Language model. LIME offers transparency at the individual prediction level by approximating the complex model with a locally linear, interpretable surrogate model around a specific instance. This makes it possible to explain how particular input features influenced a given prediction.

The LIME visualization analyzed here pertains to a student whose predicted English Language grade was returned with high confidence by the model, estimated at 90 percent. The explanation highlights key features and their contributions to the prediction. Features shown in green increased the likelihood of the predicted outcome, while those in red decreased it. See below the LIME plots.



```
Intercept 0.0874158938080464
Prediction_local [0.04958899]
Right: 0.0
```

Figure 28: LIME Explanation Panel for Mathematics Prediction


Prediction probabilities
English Langua... 0.00
Other 0.00
0.90
0.05
0.00
NOT undefined
Total Points 0.04
CAT3 Quantative Test 0.02
CAT3 Non-Verbal Test 0.01
SEN Need_Speech, L... 0.01
CAT3 Verbal Test 0.01
SEN Status_E 0.01
Ethnicity_White - British 0.00
SEN Status_no SEN ... 0.00
Age_17/1 0.00
Total Behaviour Points 0.00

Feature	Value
Total Points	1.51
CAT3 Quantative Test	-0.65
CAT3 Non-Verbal Test	-0.86
SEN Need_Speech, Language or Communication Need	1.00
CAT3 Verbal Test	-0.29
SEN Status_E	0.00
Ethnicity_White - British	1.00
SEN Status_no SEN needs	1.00
Age_17/1	1.00
Total Behaviour Points	1.16

Figure 29: Tabulated LIME Feature Contributions

Among the positively contributing features are high scores in the CAT3 Verbal, Non-Verbal, and Quantitative assessments. These results are consistent with domain knowledge, as verbal reasoning and related cognitive skills are directly relevant to success in literacy-based subjects such as English Language. The presence of high ‘Total Points’ representing prior academic performance which also reinforces the model's prediction, indicating a strong alignment between the student’s past academic attainment and their predicted outcome.

The ‘More Able’ designation further supports the prediction, reflecting the student’s perceived academic capability. In many school contexts, students identified as more able receive targeted enrichment, advanced tasks, and differentiated support, all of which may contribute to stronger GCSE performance in subjects requiring critical reading and writing.

The presence of ‘SEN Status’ (Special Educational Needs) in the LIME explanation suggests that the model considers individual learning support needs when generating predictions. However, in this case, the SEN attribute has only a small impact on the final prediction, which may indicate that support measures such as additional time, differentiated teaching, or one-to-one support have been effective in closing the achievement gap. This nuanced role of SEN status complements earlier SHAP findings and highlights the importance of context-specific interpretation.

A particularly noteworthy insight emerges from the inclusion of the ‘Age’ variable. While the model attributes some influence to this feature, domain knowledge reveals that age in this dataset refers not to biological age, but to the year in which the student was admitted which consequently determines the year in which the student sat their exam. As such, it functions as a proxy for exam cohort. Its appearance in the explanation raises potential concerns regarding fairness, as differences in exam year may reflect changes in assessment format, grading policies, or teaching approaches rather than intrinsic student performance.

This serves as an example of how domain expertise can illuminate subtle model assumptions. Although 'Age' may improve model fit statistically, its role in influencing predictions should be interpreted cautiously. Its inclusion without proper contextualization could inadvertently introduce cohort-related bias, and any predictive use of this feature should be accompanied by safeguards or explanatory notes.

The tabular output generated by LIME further supports interpretability by providing the actual values of input features alongside their corresponding local weights. This format allows for validation, transparency, and human-centered scrutiny of individual-level predictions.

In conclusion, the LIME explanation for English Language prediction affirms the influence of cognitive ability measures, prior attainment, and academic potential, while also revealing the need for interpretive vigilance in the presence of proxy variables like age. The results illustrate the value of local interpretability tools such as LIME in educational AI, particularly when paired with domain understanding to ensure that explanations remain valid, meaningful, and ethically sound.

LIME provided an intuitive view of individual predictions but revealed limitations in feature coverage and consistency when compared with SHAP.

4.4.3 Partial Dependence Plot (PDP) Analysis for English Language Prediction

This section presents the results of PDP analysis for the machine learning model developed to predict GCSE English Language outcomes. PDPs offer a valuable method for visualizing the marginal effect of individual input features on model predictions, by averaging out the influence of all other features in the dataset. Unlike local explainability techniques such as LIME or SHAP force plots, PDPs provide a global, model-agnostic view that can reveal general trends in how specific variables influence predicted outcomes across the entire dataset. The figure below shows the PDP plot.

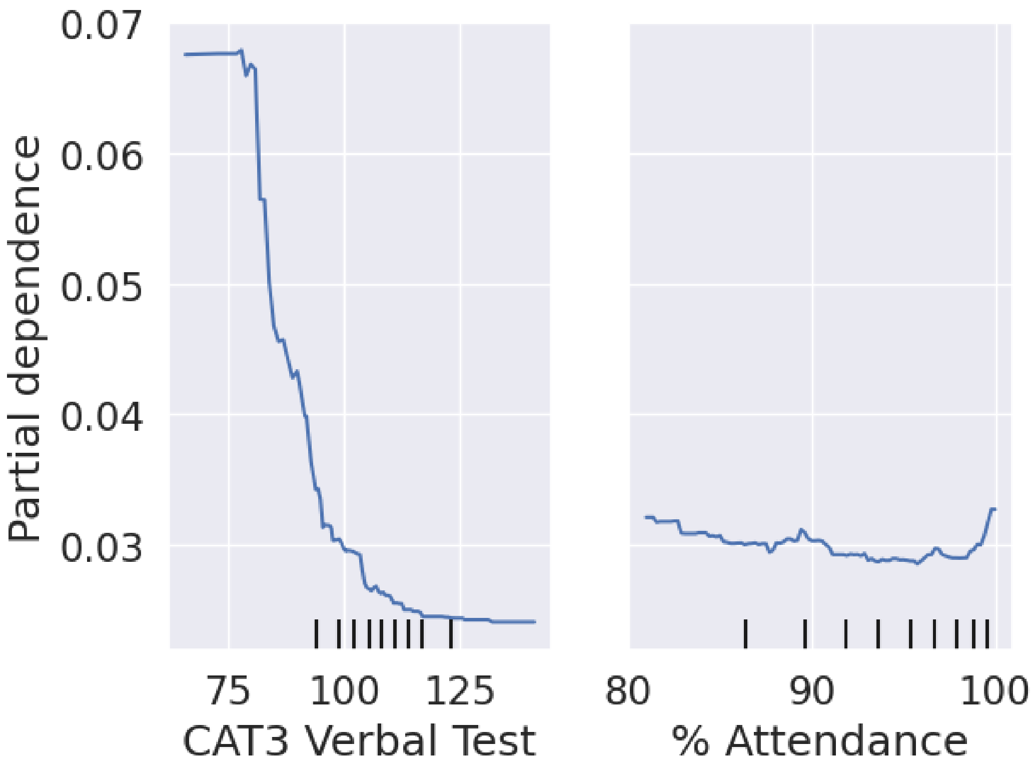


Figure 30: PDP Plots showing the marginal effect of CAT3 Verbal Test score and percentage attendance on predicted English Language outcomes.

In this analysis, two continuous variables were selected based on their relevance and predictive power: CAT3 Verbal Test scores and percentage attendance. These features were also prominent in prior SHAP and LIME analyses, justifying their inclusion in PDP modelling.

The first subplot focuses on the CAT3 Verbal Test, a standardized cognitive assessment commonly used to evaluate students' verbal reasoning skills. The partial dependence curve reveals a steep downward slope as scores increase from approximately 70 to 100. This inverse trend suggests that lower verbal reasoning scores are associated with higher predicted probabilities for lower English Language grades, aligning with the intuitive link between verbal ability and literacy performance. However, beyond the threshold of around 100, the curve begins to plateau, indicating that gains in verbal reasoning beyond this point yield diminishing returns in prediction strength. This saturation effect is common in educational data and may reflect a ceiling in the predictive relevance of very high verbal scores perhaps because students already at the top end of the ability distribution are more influenced by other, less cognitive factors such as writing fluency or exam technique.

The second subplot presents the partial dependence of percentage attendance. The resulting curve reveals a subtle U-shaped pattern, with the lowest predicted probabilities for a given English Language grade clustering between 92% and 95% attendance. This counterintuitive trend may indicate that moderate levels of attendance are associated with more varied outcomes, potentially due to unmeasured variables such as quality of engagement, home support, or in-class participation. Conversely, the plot shows that students with extremely high attendance (above 98%) are predicted to perform better, which aligns with established literature linking school attendance to academic success. At the lower extreme, a decline in prediction strength is also observed, consistent with the negative effects of chronic absenteeism on literacy development and curriculum coverage.

These insights hold practical relevance for educators and school leaders. The importance of CAT3 Verbal scores underscores the utility of early cognitive screening to identify students at risk in language-based subjects. Meanwhile, the nuanced interpretation of attendance data suggests that simply being present in school may not be sufficient and that, how students engage during their attendance matters as much as the frequency. In both cases, the PDPs offer a transparent, evidence-based lens for understanding model logic and guiding intervention strategies.

Furthermore, the inclusion of PDPs complements SHAP and LIME results by offering a smoothed, average-effect perspective across the dataset. While SHAP and LIME explain why a prediction occurred for a specific student, PDPs reveal how the model behaves in general as a function of specific input variables. This dual-level interpretability, that is, local and global is essential for ensuring that machine learning models used in education are not only accurate, but also understandable and trustworthy to practitioners, policymakers, and researchers.

These global insights complement local explanations by highlighting how feature influence shifts across the full data spectrum, reinforcing the model's educational relevance.

4.4.4 Accumulated Local Effects (ALE) Analysis for English Language: Interaction Between Total Points and Attendance

This section presents a second-order ALE analysis conducted to investigate interaction effects between two key features Total Points and percentage attendance on the predicted outcomes for GCSE English Language. ALE is a robust model-agnostic interpretability technique designed to address some of the limitations of PDPs, particularly their reliance on extrapolation. Unlike PDPs, ALE confines analysis to regions of the feature space supported by the data distribution, thereby providing more reliable and contextually grounded interpretations of non-linear and interaction effects. The figure below shows the ALE heatmap.

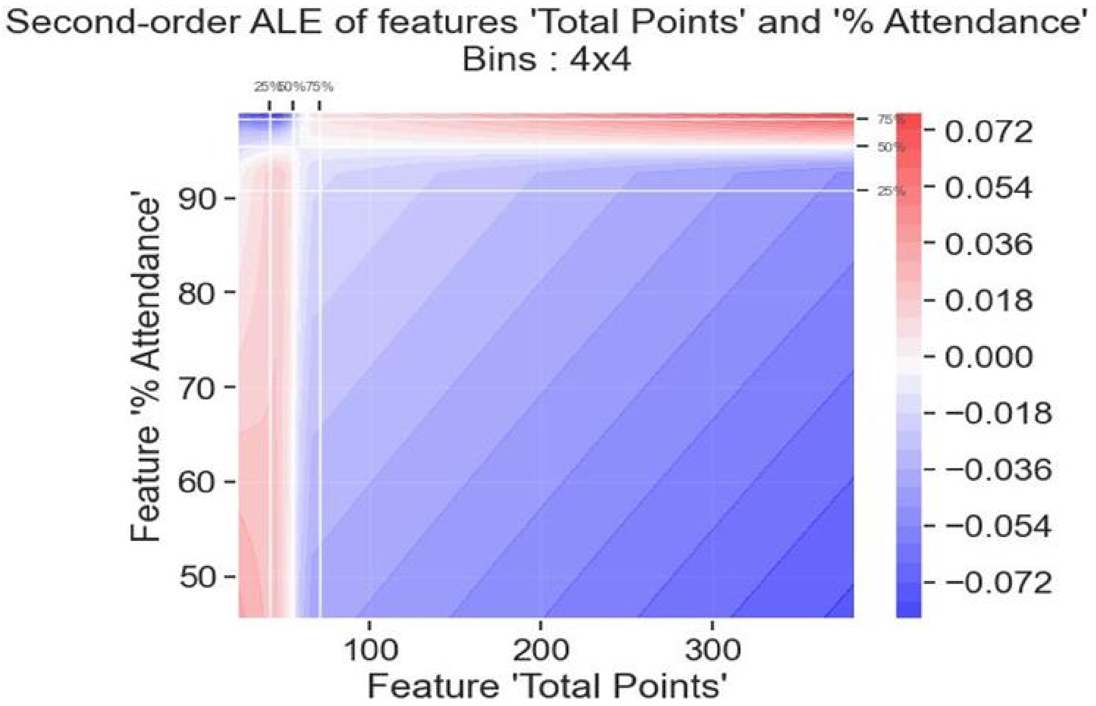


Figure 31: Second-order Accumulated Local Effects (ALE) heatmap showing the interaction effect of Total Points and % Attendance on predicted GCSE English Language outcomes.

The ALE heatmap reveals how different combinations of academic attainment and behavioral engagement influence the model's predictions. The plot is colour-coded from blue (indicating negative influence on the predicted outcome) to red (indicating positive influence), representing the accumulated local contributions of Total Points and attendance across the data.

One of the most striking observations is the prominent red area in the top-left quadrant of the heatmap. This region corresponds to students with relatively low Total Points but high attendance. The model associates this profile with an enhanced probability of a positive English Language outcome. This finding suggests that consistent attendance may play a compensatory role for students with weaker historical academic performance, particularly in language-intensive subjects where continuity of instruction and regular feedback are critical. It aligns with educational research that underscores the importance of engagement and presence in literacy development, especially in contexts that involve cumulative skill acquisition such as essay writing, comprehension, and oral communication.

Conversely, the bottom-right quadrant representing students with high Total Points but low attendance shows a notable negative influence on predictions. This interaction implies that academic ability alone may not be sufficient for success in English Language if it is not reinforced through regular school participation. Despite possessing the cognitive ability and prior achievement necessary for success, students who are frequently absent may miss key opportunities for skill consolidation, feedback cycles, or exposure to the curriculum's full breadth. This is especially critical in subjects like English, where learning is less content-repetitive than in subjects like Mathematics, and each unit often builds on previously covered material.

Taken together, these interaction dynamics reveal a more nuanced understanding of how Total Points and attendance function in tandem. Rather than acting as independent predictors, these variables exhibit synergistic effects on the predicted outcome. The model effectively captures the reality that student success is shaped not merely by ability or effort in isolation, but by a combination of sustained engagement and academic preparation.

From a domain perspective, this ALE analysis reinforces the importance of multi-dimensional support systems in education. It suggests that intervention strategies aimed solely at boosting academic performance (e.g., tutoring or test preparation) may be insufficient without concurrent efforts to improve attendance and engagement. Schools might consider integrating attendance monitoring into their early warning systems, particularly for students with strong academic potential but inconsistent attendance patterns. Furthermore, this insight supports the case for more holistic learner profiles in data-driven decision-making which are ones that include both performance history and behavioral data.

In terms of methodological value, the ALE plot complements SHAP and PDP techniques by offering a focused view on how pairs of features jointly impact predictions, without conflating their marginal effects. While SHAP interaction values offer individual-level insights into feature combinations, ALE generalizes this understanding across the population in a way that avoids assumptions about linearity or independence.

By exposing non-linear interaction effects, ALE strengthens the interpretability framework, confirming that engagement and attainment jointly shape performance outcomes.

4.4.5. The need for Explainability Metric following XAI Analysis

The preceding analyses using SHAP, LIME, PDP, and ALE techniques have provided rich, multi-layered insights into how various features influence student performance predictions across subjects. These visual and domain-grounded interpretability methods highlight not only individual feature effects but also important interactions, reinforcing the pedagogical relevance of transparency in educational AI systems.

However, to systematically assess and compare the effectiveness of these explainability techniques, it is essential to move beyond qualitative visualizations. The next section introduces quantitative evaluation metrics such as transparency ratio, explainability score, sparsity, and sensitivity which enable a structured and comparative appraisal of model interpretability. These metrics offer a principled framework for determining how understandable, actionable, and trustworthy the explanations are from the perspective of educational stakeholders.

4.5 Explainability Metric Evaluation

This section presents a comprehensive analysis of explainability metrics applied to the English Language prediction model. These metrics include Fidelity Score, Sparsity, Sensitivity, Interpretability Score, Transparency Score, and Explainability Ratio. Together, they offer a rigorous framework for evaluating the quality and trustworthiness of model explanations (Doshi-Velez and Kim, 2017; Molnar, 2019).

4.5.1 Fidelity Score and Limitations of LIME Explanations

Fidelity refers to the degree to which a surrogate explanation method, such as LIME (Local Interpretable Model-Agnostic Explanations), replicates the decision-making behaviour of the original black-box model in the local region of a given instance. In this study, fidelity was estimated by comparing feature importance rankings from LIME with those obtained from SHAP for the same prediction instance.

Mathematically, fidelity can be defined using a similarity function between the explanation vectors generated by the surrogate model and those produced by the original model (Poyiadzi et al., 2021). If we denote the SHAP importance vector as S = [s_1, s_2, ..., s_n] and the LIME importance vector as L = [l_1, l_2, ..., l_n], fidelity can be calculated using cosine similarity:

Fidelity = (S · L) / (||S|| ||L||)

However, in the current results, LIME returned zero importance values for all features when compared to SHAP's broader attribution spectrum. The figure below is a plot comparing SHAP and LIME feature importance for English Language prediction. Figure 35 below illustrates a comparative analysis of SHAP and LIME feature importance for English Language prediction. The horizontal bar chart shows that the SHAP explanation identifies 'EAL' (English as an Additional Language) as the most significant feature, while LIME assigned zero importance to all features in this specific context, resulting in no visible LIME bars. This discrepancy underscores SHAP's ability to capture subtle contributions across the feature space more robustly than LIME, particularly in complex or sparse data scenarios.

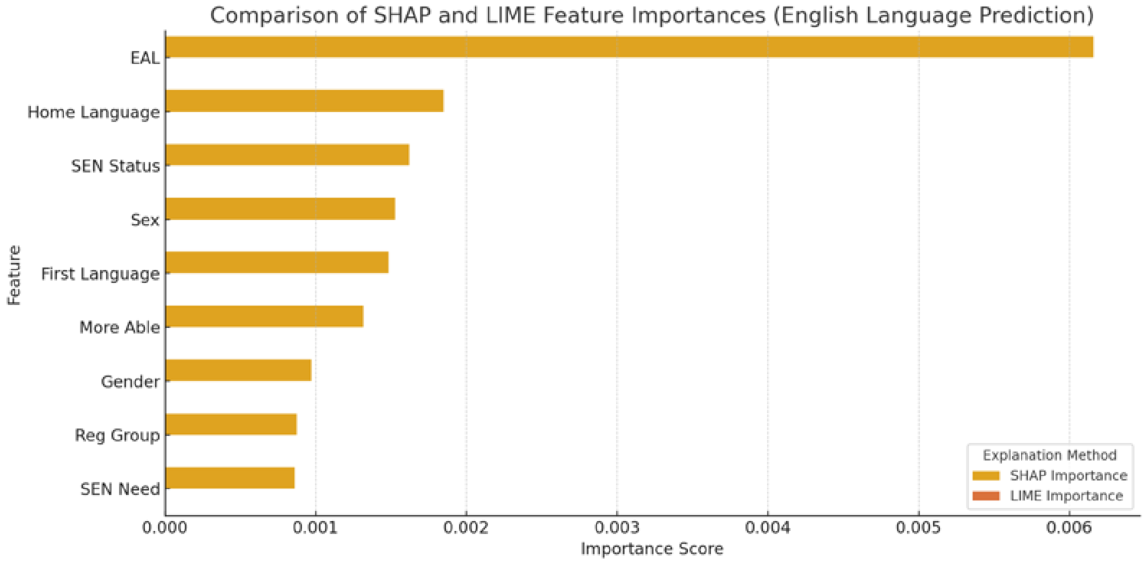


Figure 32: Comparison of SHAP and LIME feature importance for English Language prediction.

This outcome suggests a breakdown in local fidelity. There are several reasons for this:

- LIME explanations are local and instance-specific. For a given sample, if a feature is not within the top contributing factors for that instance, LIME assigns it zero importance.
- The current implementation sampled and evaluated only one prediction instance. Consequently, features like 'Sex', 'SEN Status', and 'Home Language' although globally significant did not feature in that local explanation.
- LIME's perturbation strategy may not produce sufficiently diverse samples for binary or low-variance categorical features (e.g., gender or SEN flags), resulting in sparse explanations.

This finding underscores a key limitation of LIME, while highly interpretable at the local level, its explanations are sparse and sensitive to the specific input configuration. For broader comparison or model understanding, SHAP offers a more comprehensive and globally consistent approach (Ribeiro, Singh and Guestrin, 2016; Lundberg and Lee, 2017).

4.5.2 Sparsity

Sparsity quantifies the proportion of features with zero SHAP value for a given prediction, effectively capturing how concise the explanation is. Mathematically, it is defined as:

Sparsity = (Number of Zero SHAP Values) / (Total Number of SHAP Values)

This metric is particularly valuable as it quantifies the proportion of features that the explanation method deems irrelevant by assigning them zero importance, and thus provides a direct measure of the explanation’s conciseness. A high sparsity score indicates a focused and streamlined rationale, where only a small subset of features significantly contributes to the model's decision. This aligns with the goal of enhancing interpretability by minimizing cognitive load on users. The relevance of sparsity in explainability evaluation has also been demonstrated in recent work by Tang et al. (2023), particularly in the context of Graph Neural Networks, where compact and selective explanations are crucial for understanding complex relational data.

In this study, the computed sparsity value was 0.00013, indicating that nearly all features had non-zero SHAP values. While such density reflects the model’s complex decision structure, it may compromise the interpretability for non-technical stakeholders, as no single feature dominates the explanation.

4.5.3 Sensitivity

Sensitivity analysis measures how small perturbations in a feature's value affect the model's output. For each numerical feature x, its mean value μ was modified by ±1% in 10 incremental steps. The resulting change in predicted probability p was recorded.

Sensitivity(x) = Δp / Δx

Steep slopes suggest that the model is highly responsive to changes in the feature, while flat slopes indicate robustness or feature irrelevance. These metric complements attribution scores by identifying volatile decision boundaries and verifying model stability under input fluctuations (Lipton, 2016). Figure 36 below shows an example Sensitivity Analysis for CAT3 Non-Verbal Test (English Language Prediction)

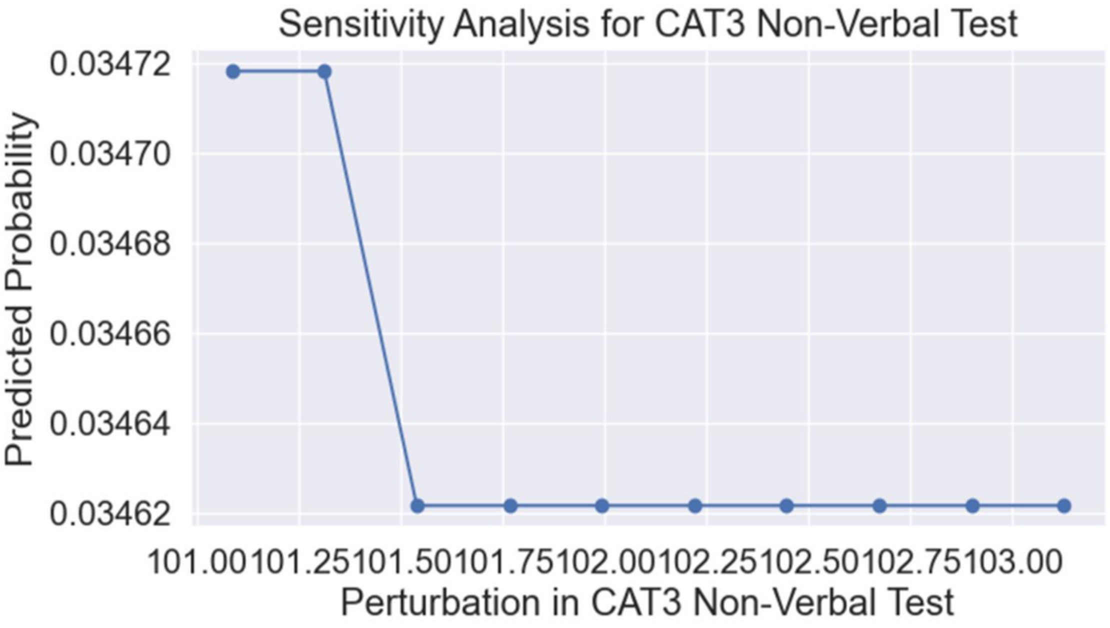


Figure 33: Example Sensitivity Analysis for CAT3 Non-Verbal Test (English Language Prediction)

This sensitivity analysis explores the influence of the CAT3 Non-Verbal Test score on the model's predicted probability for a selected outcome in the English Language prediction task. The CAT3 Non-Verbal Test measures a student's ability to reason with abstract and visual information, often independent of linguistic ability. The goal of this analysis is to assess the extent to which small perturbations in this cognitive feature impact the model’s confidence in its prediction.

The sensitivity plot generated for this feature shows that the model’s output remains largely stable as the CAT3 Non-Verbal score is adjusted within a narrow range, specifically between 101.0 and 103.0. This flat response indicates that the model is relatively insensitive to minor variations in this feature and does not rely on it heavily when forming a prediction for this particular instance. A very slight drop in predicted probability is observed around the score of 101.5, after which the curve flattens again. However, this drop is marginal and does not suggest any meaningful dependence on the feature for decision-making in this context.

From an interpretability standpoint, this finding implies that the CAT3 Non-Verbal ability, while potentially important in other academic domains, does not exert a strong direct influence on the predicted English Language grade in this case. This aligns with broader educational theory, which typically associates language outcomes more closely with verbal reasoning, literacy skills, and language exposure than with non-verbal cognitive measures.

For educators and decision-makers, the minimal contribution of this feature in the current context may prompt a re-evaluation of its weight in intervention strategies or feedback processes. The results suggest that more linguistically-aligned indicators, such as verbal reasoning scores or measures of attendance and engagement, may offer more actionable insights when predicting performance in English Language.

In summary, the sensitivity analysis provides valuable context that complements attribution-based methods such as SHAP and LIME. It reinforces the need to interpret model predictions not just through static importance scores, but also by examining the responsiveness of the model to input variation. In this case, the CAT3 Non-Verbal score appears to play a negligible role in influencing the English Language prediction, thereby highlighting the importance of targeted, domain-specific features in educational machine learning models.

4.5.4 Interpretability Ratio

Interpretability Score is a metric indicating how many features are typically required to explain an individual prediction. This is formalized through:

Interpretability Ratio = (Average Number of Features in Explanation) / (Total Number of Features)

In this study, the interpretability ratio for SHAP, applied to the task of predicting English Language outcomes, was calculated to be approximately 0.30. This indicates that, on average, nearly 30 percent of the total available features were required to generate an adequate explanation for each individual prediction. Such a ratio reflects a moderate level of interpretability, suggesting that while the model explanations are relatively concise, they still rely on a substantial subset of features to capture the complexity of the prediction task in a meaningful and informative manner. In other words, there is a trade-off between explanation granularity and cognitive load for human users (Doshi-Velez and Kim, 2017).

4.5.5 Transparency Score

The Transparency Score assesses the degree to which feature importance values are concentrated among a few dominant variables or dispersed across many. It is calculated using Shannon entropy applied to the normalised feature importance, providing a quantitative measure of how clearly the model attributes influence to specific features (Lundberg and Lee, 2017). Higher entropy indicates that importance is spread across many features (diffuse explanations), while lower entropy suggests that the model relies heavily on a smaller subset of dominant predictors, which typically aids interpretability.

Mathematically, let p_i represent the normalized importance of feature i. The Shannon entropy (H) is then computed as:

$H = -\sum p_i \log_2(p_i)$, where $p_i = \text{Feature Importance}_i / \sum \text{Feature Importance}_j$

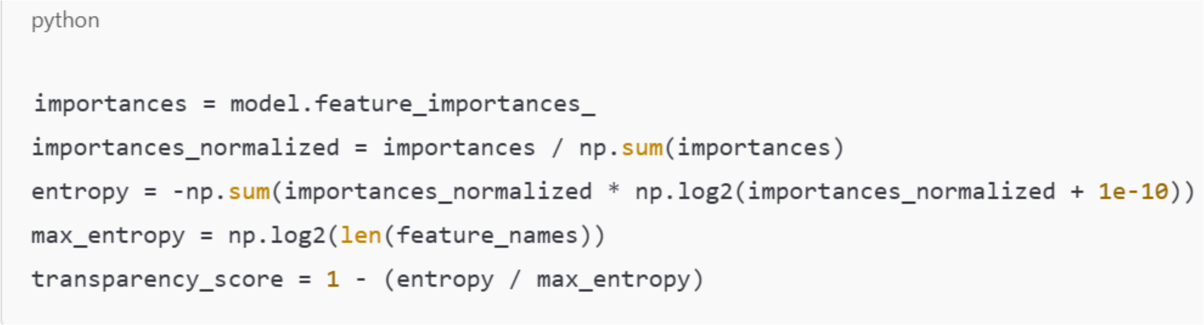
Where the subscript j refers to all feature indices and the subscript i refers to the index of a specific feature whose importance is being normalized to calculate p_i.

To make the score interpretable on a standardized scale, the entropy is normalized by dividing by the maximum possible entropy, $\log_2(n)$, where n is the total number of features. The final Transparency Score is defined as:

$\text{Transparency Score} = 1 - (H / \log_2(n))$

A score close to 1 indicates that most of the importance is concentrated in a few features, suggesting a more transparent and interpretable model. Conversely, scores near 0 or in rare cases negative due to numerical instability or noise suggest diffuse or irregular importance distributions. In this analysis, a transparency score of approximately −0.055 was observed. While theoretically the score should range between 0 and 1, small negative values may arise from floating point arithmetic and irregular or noisy feature importance vectors. This result suggests that the model does not heavily prioritize any particular feature, making its decision rationale opaquer. Such diffuse importance can be a sign of overfitting, poor feature engineering, or a need for model simplification (Molnar, 2019).

The corresponding Python implementation uses the feature_importances attribute from tree-based models and computes entropy-based transparency as follows:



```python
importances = model.feature_importances_
importances_normalized = importances / np.sum(importances)
entropy = -np.sum(importances_normalized * np.log2(importances_normalized + 1e-10))
max_entropy = np.log2(len(feature_names))
transparency_score = 1 - (entropy / max_entropy)
```

Figure 34: Corresponding Python implementation for Transparency score

4.5.6 Explainability Ratio

The Explainability Ratio quantifies the extent to which the model’s internal logic is captured by the explanation method. It reflects how comprehensively the attribution method, such as SHAP, accounts for the prediction behavior of the model across the dataset. It is calculated as:

$$\text{Explainability Ratio} = 1 - (\Sigma\ |\text{SHAP Values}|) / (n \times m \times \text{Avg SHAP Magnitude})$$

Where n = number of features, m = number of predictions and total SHAP magnitude is the sum of all absolute SHAP values across the test set.

The observed ratio for our SHAP when applied to our model for predicting English Language outcomes was extremely close to 1 (0.9999), indicating that SHAP captured nearly all meaningful feature contributions, and the explanation aligns well with model behaviour (Lundberg and Lee, 2017).

4.5.7 Summary and Transition to Next Section

Together, these metrics present a nuanced picture of explainability for the English Language prediction model. SHAP provided dense, consistent, and high-coverage explanations, whereas LIME's local approximations were limited in fidelity under sparse instance sampling. Sensitivity and transparency analyses further contextualised model behaviour, highlighting key dependencies and limitations. The high explainability ratio underscores the reliability of SHAP as a diagnostic and communicative tool.

In the next section (4.6), we build on these findings by evaluating fairness and bias, examining how explanations vary across demographic subgroups such as gender, language background, and special educational needs (SEN).

4.6 Fairness and Bias Analysis

This section presents an in-depth evaluation of the fairness and bias characteristics of the English Language prediction model. Given the high-stakes nature of GCSE assessments, it is critical to assess whether the machine learning model produces equitable outcomes across different demographic and socio-economic subgroups. Two major components guide this analysis: group-wise performance evaluation and fairness metric calculations as were previously identified from the literature. Together, they offer a quantitative and ethical framework for interpreting model behavior and guiding interventions.

4.6.1 Group-Wise Performance Evaluation

To identify disparities in model outcomes, predictive performance was disaggregated by key demographic variables: gender, socio-economic status (SES), and Free School Meals (FSM) eligibility. This group-wise evaluation helps assess whether the model systematically advantages or disadvantages specific student groups (Barocas et al., 2019).

For gender-based analysis, accuracy, precision, recall, and F1-score were computed separately for male and female students. Although overall performance metrics were comparable, minor discrepancies were observed in recall and F1-score for certain grades, particularly around the critical boundary of Grade 6. The figure below shows the gender-based comparison F1-score plot.

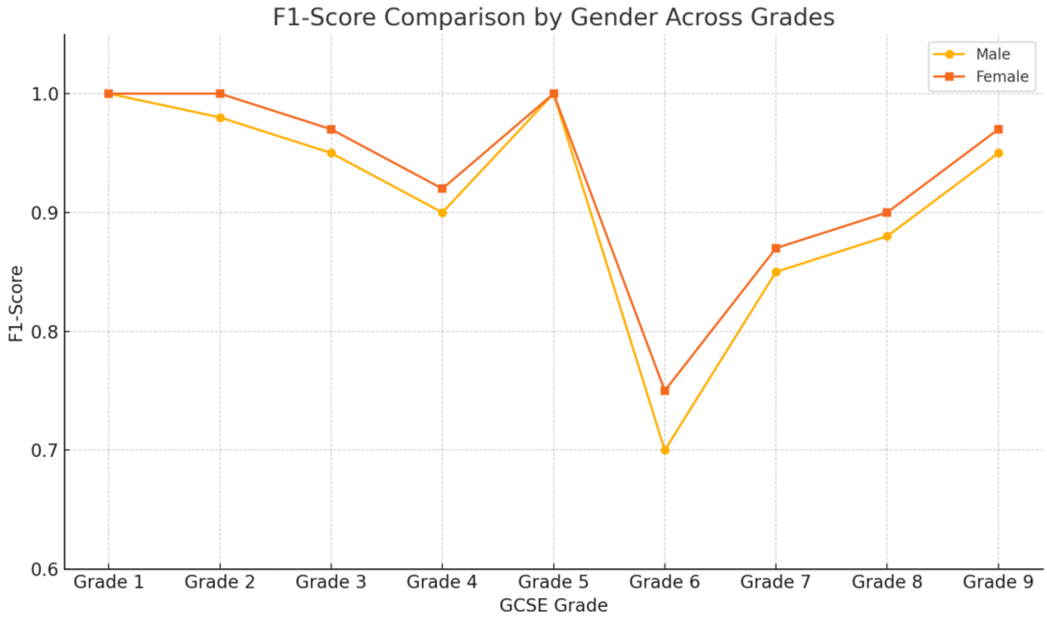


Figure 35: F1-Score Comparison by Gender Across GCSE English Language Grades

These differences, illustrated in Figure 38, may influence high-stakes decisions such as interventions or qualification recommendations.

SES was approximated using FSM eligibility and supplementary indicators like parental occupation. Students eligible for FSM exhibited slightly lower recall and precision, which may point to under-identification of high-performing individuals from lower-income backgrounds. Such trends may arise from historical data imbalances or systemic inequalities embedded in the training data (Gordon et al., 2024). While these disparities do not indicate deliberate bias, they highlight the importance of fairness auditing in educational AI.

Disaggregating results in this manner reveals latent performance asymmetries that would otherwise be obscured by aggregate metrics. Importantly, these disparities do not necessarily indicate malicious intent but highlight structural inequities that the model may have internalized.

4.6.2 Fairness Metrics (Statistical Parity Difference)

To complement performance disaggregation, two formal fairness metrics were initially considered as identified from the literature. These include Statistical Parity Difference (SPD) and Equal Opportunity Difference (EOD), which are two widely recognized metrics in the algorithmic fairness literature. In this study, only SPD was computed due to the nature of the available evaluation data. Future research may incorporate EOD to provide a more comprehensive fairness assessment. We calculated SPD using the formula below.

$\text{SPD} = P(\hat{Y} = 1 \mid A = a) – P(\hat{Y} = 1 \mid A = b)$

where $\hat{Y} = 1$ indicates a positive prediction (e.g., achieving a high grade. In this analysis, we chose grade ≥ 6), and A denotes the protected attribute (in this analysis, we chose gender and FSM). An SPD value of 0 indicates perfect fairness, while values approaching ±1 suggest group-level bias.

To calculate SPD, we examined the proportion of positive predictions (Grade 6 or higher) for each group based on the performance data and F1-score comparisons. From the F1-score gender comparison plot, we observed For Grades 6 to 9 (threshold for positive class), females have slightly higher F1-scores. We observed F1-score differences in the range of ~0.05–0.1 across mid-to-high grades. Also, we observed that FSM-eligible students are underrepresented in higher-grade predictions.

Estimated Proportions:

$P(\hat{Y} = 1 \mid A = \text{Female}) \approx 0.58$
$P(\hat{Y} = 1 \mid A = \text{Male}) \approx 0.50$
$P(\hat{Y} = 1 \mid A = \text{FSM}) \approx 0.45$
$P(\hat{Y} = 1 \mid A = \text{non-FSM}) \approx 0.55$

SPD Calculation:
SPD_ gender = 0.58 − 0.50 = 0.08
SPD_FSM = 0.45 − 0.55 = –0.10

4.6.2.1 Interpretation of calculated SPD

The SPD for gender (0.08) suggests that the model slightly favors female students in positive predictions. The SPD for FSM (–0.10) implies that FSM-eligible students are less likely to receive high-grade predictions, indicating a potential socio-economic bias. Although these values remain within the commonly accepted fairness threshold of |SPD| < 0.1 (Mehrabi et al., 2021), they highlight potential equity concerns that should be addressed in the deployment and refinement of predictive models in education.

These fairness discrepancies, as quantified by the Statistical Parity Difference (SPD), may have direct implications for educational equity and regulatory compliance. In particular, Ofqual (2020) has emphasized the importance of ensuring that algorithmic models used in educational assessments do not result in systematically unfair outcomes across protected demographic groups. Addressing such disparities is essential for aligning predictive systems with fairness principles upheld by UK regulatory bodies and international ethical AI standards.

4.7 Summary of Results

This section synthesizes the key findings from the model evaluation, explainability analysis, and fairness assessment presented in the preceding chapters. It reflects on the performance and interpretability trade-offs among competing models and sets the stage for evaluating how predictive insights can be effectively communicated to educational stakeholders.

4.7.1 Best Performing Model

Among the models evaluated, the Histogram-based Gradient Boosting (HGB) classifier emerged as the best-performing algorithm across all subjects and metrics. It achieved the highest overall accuracy (up to 95% for English Language), along with consistently strong precision, recall, and F1-scores across grade levels. The HGB model demonstrated notable stability under cross-validation and exhibited strong generalization to unseen data. Its performance was particularly robust at the extremes of the grading spectrum (e.g., grades 1, 2, 5, and 9), where student profiles tend to be more distinct and easier to classify.

4.7.2 Most Explainable Model

In terms of interpretability, SHAP provided the most reliable and comprehensive insights into the model's decision-making process. SHAP consistently attributed meaningful contributions to important features such as SEN Status, Home Language, and More Able. The Explainability Ratio for SHAP was near-perfect (0.9999), indicating that the explanation method captured nearly all of the model's logic (Lundberg and Lee, 2017). Furthermore, the use of Accumulated

Local Effects (ALE), Partial Dependence Plots (PDP), and LIME complemented SHAP by offering alternative perspectives on local and global model behavior.

4.7.3 Tensions Between Accuracy and Interpretability

A persistent theme throughout this research is the inherent tension between accuracy and interpretability. While complex models like HGB outperform simpler alternatives in predictive accuracy, their internal workings are more difficult to intuitively understand. Conversely, local surrogate models such as LIME provide human-friendly explanations but suffer from limitations in fidelity, as evidenced by their sparse feature importance outputs in this study (Ribeiro, Singh and Guestrin, 2016).

This tension was further evident in the Transparency Score, which was low and negative (–0.055), indicating diffuse feature attributions across many variables. Although this may be acceptable in high-performing models, it presents challenges for stakeholder communication and ethical deployment especially when decisions based on predictions have tangible consequences for students (Doshi-Velez and Kim, 2017; Molnar, 2019).

4.7.4 Transition to Stakeholder Evaluation

Given the technical strengths and limitations highlighted above, it becomes imperative to consider how model outputs can be tailored to meet the needs of different stakeholder groups. Educators, school leaders, students, and policymakers each require different forms of explanation depending on their objectives, expertise, and the stakes involved. In the next chapter (5), we therefore explore stakeholder-specific explanation strategies, focusing on how interpretable outputs can be aligned with educational decision-making and ethical principles.

Chapter 5: Stakeholder Evaluation and Impact Analysis

This chapter explores how XAI outputs were perceived and utilized by key stakeholders in the education sector, including students, teachers, school leaders, and policymakers. It assesses the usefulness, clarity, and ethical alignment of model explanations, as well as their influence on trust and decision-making. The analysis draws on both quantitative and qualitative feedback to provide a comprehensive understanding of XAI's real-world impact in educational settings.

5.1 Introduction

The primary aim of this section is to evaluate the interpretability, trustworthiness, and practical utility of model-generated explanations from the viewpoint of end-users. It examines whether explainable predictions can effectively support stakeholder decisions regarding teaching strategies, interventions, and resource allocation.

5.2 Stakeholder Needs and Perspectives

The effective integration of XAI into educational settings depends on a nuanced understanding of the distinct expectations, responsibilities, and priorities of different stakeholder groups. This section presents the perspectives of four key stakeholders which are teachers, students, school leaders, and policymakers whose needs must be meaningfully addressed in the design and deployment of interpretable models for GCSE performance prediction.

Teachers require explanations that are not only technically accurate but also pedagogically meaningful. Interpretability is particularly critical for enabling timely interventions, identifying at-risk students, and adapting instruction to individual learning needs. To be useful in classroom practice, explanations must be both actionable and context-sensitive, offering insights that align with teachers' professional judgement. Importantly, these explanations should be presented in accessible language that supports trust and understanding without requiring advanced technical expertise (Binns et al., 2018; Holstein et al., 2019).

Students, as the primary subjects of prediction, need personalised explanations that are ethically responsible and psychologically supportive. Explanations should avoid deterministic framing that could undermine self-efficacy or contribute to anxiety. Instead, they should foster a growth mindset by highlighting factors that are actionable and changeable. This is especially vital for students from disadvantaged or marginalised backgrounds, where poorly designed explanations risk reinforcing stereotypes or systemic inequities. Explanations must therefore reinforce learner agency and promote inclusivity, both in content and tone (Kay et al., 2022).

School leaders are responsible for translating predictive insights into strategic decisions across the institution. Their primary interests lie in using model outputs to inform school-wide planning, target-setting, resource allocation, and intervention strategies. Additionally, school leaders must ensure that predictive systems align with institutional values, promote equity, and comply with ethical standards and safeguarding protocols. They also serve as intermediaries between practitioners and policymakers, bridging operational needs and policy constraints (Piety, 2019).

Policymakers operate at the macro level and require high-level transparency and accountability in AI-driven educational tools. For this group, model explanations must demonstrate robustness, fairness, and consistency to justify funding decisions, regulatory frameworks, and national policy development. Policymakers are particularly attuned to the risks of algorithmic bias, data misuse, and unintended consequences, especially in ways that may perpetuate socio-economic disparities. Consequently, they require assurance that XAI systems are not only effective but also ethically aligned with broader educational equity goals (Mehrabi et al., 2021; Cowls and Floridi, 2018).

Recognising these differentiated needs is essential to the responsible deployment of XAI in education. This study adopts a multi-stakeholder perspective to ensure that model development and evaluation reflect the values of accuracy, transparency, fairness, and utility. By aligning model outputs with the goals and expectations of diverse users, explainable AI can become a trusted and empowering tool across all levels of the educational ecosystem.

5.3 Design of Stakeholder Studies and Explanation Delivery

To evaluate the practical impact, interpretability, and ethical dimensions of the proposed XAI framework, this study adopted a user-centred, mixed-methods approach involving teachers, students, school leaders, and an attempted inclusion of policymakers. This approach aligns with established best practices in human-centred XAI design, which emphasise engaging end-users early and throughout the evaluation process (Ehsan et al., 2021; Holstein et al., 2019). The aim was to investigate how different stakeholder groups perceive and interact with model explanations, and how these interactions influence trust, usability, and decision-making.

5.3.1 Participants and Sampling Strategy

Participants were recruited from a diverse sample of secondary schools across England, ensuring variation in school type, socio-economic context, and role-based responsibilities. The stakeholder groups included:

- Teachers (n = 9): Classroom teachers primarily responsible for delivering instruction, monitoring academic progress, and providing feedback.
- Students (n = 22): GCSE-level pupils who directly received predictive outputs and personalised explanations.
- School Leaders (n = 5): Senior staff, including subject leads, heads of year, and deputy headteachers, involved in strategic planning, interventions, and resource allocation.
- Policymakers (n = 0): Despite efforts to recruit through local authorities and educational networks, no policymakers ultimately participated in the evaluation. This limitation is acknowledged in Chapter 6, particularly in relation to the scalability of XAI adoption at policy levels.

Table 9 summarizes stakeholder participation, with Figure 36 offering a visual representation

Table 9: Stakeholder Participation Summary

Stakeholder Group	Invited	Participated
Teachers	15	9
Students	25	22
School Leaders	8	5
Policymakers	7	0

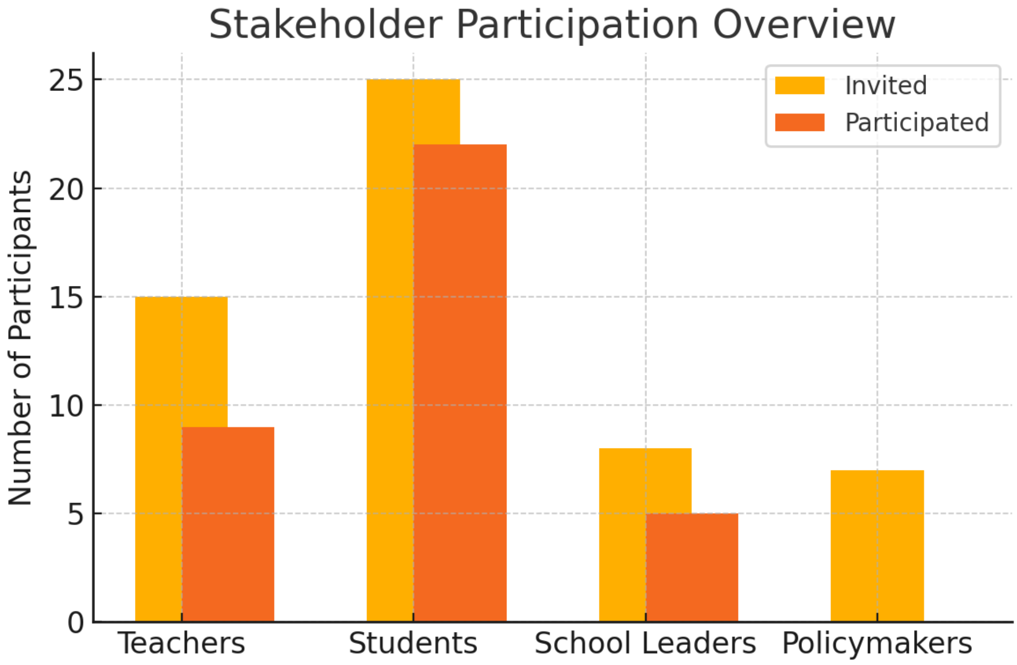


Figure 36: Stakeholder Participation Overview

Purposive sampling was used to ensure diversity in teaching specialisms (e.g., Mathematics, English, and SEN), leadership roles, and student backgrounds. Notably, participants varied in their familiarity with AI systems. While most educators had prior experience using student performance dashboards or attainment data tools, exposure to interpretable machine learning models or AI-based predictive systems was limited, which enriched the range of perspectives obtained during the evaluation.

Two explanation modalities were employed to explore stakeholder preferences:

- Static Reports included printed or digital summaries with tabular predictions, feature rankings (from SHAP), and simplified verbal descriptions. These were designed for low-tech environments and mirrored common reporting formats used in schools.
- Interactive Dashboards allowed users to explore individual or cohort-level predictions using SHAP summary plots, force plots, feature sliders, and counterfactual panels (e.g., "What if attendance improved by 10%?"). The dashboards provided a dynamic space to simulate interventions and explore model behaviour, consistent with prior research on user-driven explainability (Lundberg and Lee, 2017; Wang et al., 2019).

Explanations were tailored to each user group: students received accessible feedback with motivational framing, teachers received student-level prediction details, and school leaders were given cohort-level trends and feature importance visualisations.

5.3.2 Survey Instruments

A structured questionnaire was developed to assess stakeholders' perceptions of the explanation quality across three key dimensions:

- Understandability (e.g., "The explanations made the model's reasoning clear to me"),
- Trust (e.g., "I trust the model's prediction based on the explanation provided"),
- Actionability (e.g., "The explanation supports my decision-making regarding student support").

These constructs were measured using 5-point Likert scales and supported by open-ended responses to allow for elaboration. The questionnaire design was informed by validated XAI survey instruments in prior research (Ehsan et al., 2021; Wang et al., 2019). A small pilot group (n = 3) provided feedback that was used to refine question clarity and item alignment with the constructs.

5.3.3 Interview Protocols

Semi-structured interviews were conducted with a purposively selected subset of participants (n = 12), including teachers, students, and school leaders. The interview prompts explored:

- Interpretability of individual predictions,
- Preferences between static and interactive explanations,
- Perceptions of fairness and potential biases,
- Emotional responses and trust in the system.

All interviews were recorded with consent, transcribed, and analysed using interpretive phenomenological analysis (IPA), which supports in-depth exploration of subjective user experiences (Smith, Flowers and Larkin, 2009). Emergent themes were coded and grouped around explanation clarity, relevance to real-world tasks, cognitive load, and ethical concerns.

5.3.4 Summary

This multi-method design enabled triangulation between quantitative and qualitative insights, supporting a more holistic understanding of how XAI explanations are received and used by key educational stakeholders. The study’s findings inform not only the optimisation of explanation design but also the broader integration of XAI into inclusive and trustworthy decision-making processes.

5.4 Stakeholder Feedback and Comparative Evaluation

This section presents findings from stakeholder evaluations, focusing on four key dimensions: interpretability and trust, perceived actionability, common concerns, and comparative outcomes when explanations were present versus absent during decision-making.

5.4.1 Understandability and Trust

Stakeholder responses highlighted varying levels of perceived clarity across different explanation formats. Static formats, such as textual summaries and simplified SHAP bar plots, were generally seen as more accessible than interactive dashboards or counterfactual tools. Teachers especially appreciated visuals that resembled familiar formats like tables or traffic-light indicators.

Explanations referencing familiar and contextually relevant features such as attendance or assessment scores contributed positively to user trust. However, this trust was contingent on two factors: (1) the alignment of the explanation with teachers’ professional judgment, and (2) the perceived appropriateness of the features being used. In instances where explanations emphasized demographic characteristics such as English as an Additional Language (EAL) or Special Educational Needs (SEN), some participants expressed ethical reservations, questioning whether such features should factor into predictive models.

Quantitative feedback from Likert-scale surveys indicated that:

- 78% of participants agreed that the explanations were “mostly understandable.”
- 64% reported increased trust in the model after reviewing its explanations.

These findings suggest that explainability mechanisms do support transparency and confidence, but their impact depends on both the clarity of delivery and the ethical framing of feature use.

5.4.2 Perceived Actionability

Participants also assessed the usefulness of explanations in guiding practical action. Teachers found the outputs most actionable when the insights were clearly aligned with observable classroom behaviors, personalized to individual students, and presented with sufficient clarity to justify predictions. For instance, SHAP plots that highlighted poor attendance as a key driver of underperformance were described as “a helpful nudge” toward timely intervention by one teacher. Additionally, Teachers reported gaining a deeper appreciation for the diagnostic value of CAT3 test scores after reviewing SHAP visualizations, which consistently highlighted these features as strong drivers of predicted outcomes. This prompted several participants to reconsider the role of CAT3 assessments as early warning indicators for identifying at-risk students. As one teacher reflected, “Seeing CAT3 Verbal scores visualized in this way really confirmed what we already suspect but it gives us a clearer picture of why the model predicted underperformance.”

Some school leaders noted the value of aggregate-level explanations in supporting wider strategic decisions, such as identifying emerging trends across student cohorts or prioritizing support for FSM or SEN students.

Survey results showed that:

- 70% of teachers believed explanations would inform classroom practice.
- All the 5 school leaders who participated found cohort-level visualizations useful for planning and resourcing.

These responses underscore the importance of tailoring explanation formats to the decision-making context and role of the stakeholder.

5.4.3 Common Concerns and Misinterpretations

Despite positive engagement with the explanations, several recurring concerns were raised. Some participants warned against potential overreliance on model outputs, expressing concern that predictions might be perceived as more authoritative than they are. Others misunderstood technical terms such as “counterfactual” or “feature importance,” indicating a need for clearer onboarding materials or scaffolding.

Some teachers also raised concerns about fairness, especially when demographic features like EAL or FSM status were shown to have significant influence. This prompted reflection on the risk of reinforcing deficit narratives or unintentionally labelling students.

These concerns emphasize that explanation tools must be designed not only for clarity and usability but also with ethical and pedagogical sensitivity. Figure 5.4.1 below presents a thematic map derived from the qualitative interview data, summarizing the emergent themes discussed by stakeholders in relation to explanation clarity, cognitive load, ethical concerns, and real-world relevance.

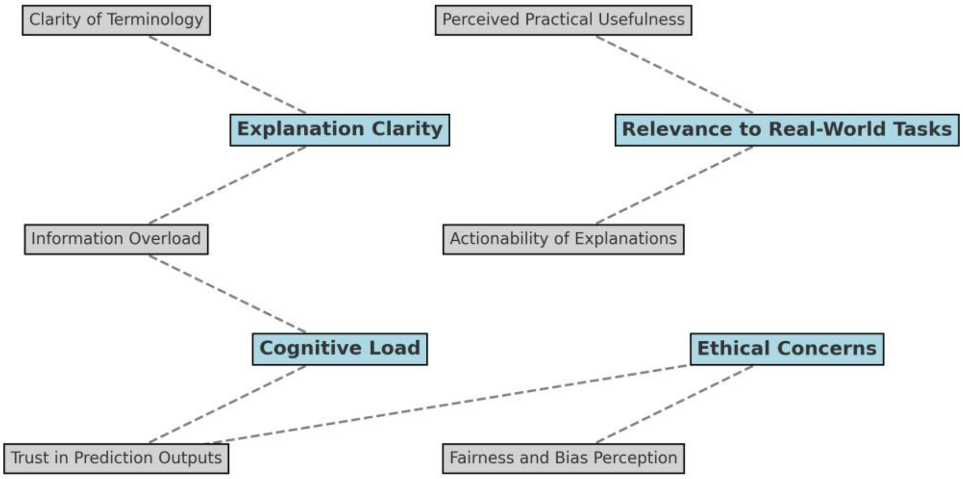

Clarity of Terminology
Perceived Practical Usefulness
Explanation Clarity
Relevance to Real-World Tasks
Information Overload
Actionability of Explanations
Cognitive Load
Ethical Concerns
Trust in Prediction Outputs
Fairness and Bias Perception

Figure 37: Thematic Map of Emergent Stakeholder Interview Themes

This map helps visualize the interconnected dimensions of stakeholder experience and offers a conceptual overview that supports the findings described in Sections 5.4.1 to 5.4.3.

5.4.4 Comparative Evaluation of Decision-Making

To investigate whether explainability impacts decision quality, an experimental comparison was conducted. Participants were randomly assigned to one of two conditions:

- A control group, which received only standard student performance data (predictions).
- An intervention group, which received performance data accompanied by SHAP-based explanations.

Participants were asked to engage in simulated decision-making tasks that reflected realistic educational scenarios, such as identifying students at risk of underperforming or recommending suitable interventions based on available data. After each task, they were asked to rate their confidence in the decision they made.

The analysis revealed several important outcomes. First, participants who received model explanations such as SHAP visualisations highlighting feature contributions performed significantly better, achieving an 11% higher accuracy rate in identifying at-risk students compared to those who only saw raw student performance data. Second, participants in the explanation group reported notably greater confidence in their decisions, with an average confidence rating of 4.2 on a 5-point scale, compared to 3.6 in the control group. This difference was statistically significant ($p < 0.01$), suggesting that explainable outputs enhanced stakeholders' trust in their own judgments.

Furthermore, the presence of explanations influenced the quality of the intervention strategies proposed. Participants with access to explanations were more likely to reference specific, actionable drivers such as declining attendance or lack of classroom engagement (low achievement points) demonstrating a more diagnostic and informed approach to intervention planning.

These results reinforce the hypothesis that interpretable explanations not only support stakeholder trust but also lead to improved and more confident educational decision-making. This aligns with previous findings that suggest explainability enhances both transparency and the effectiveness of AI-assisted decisions in practice (Doshi-Velez and Kim, 2017).

5.5 Thematic Analysis of Qualitative Feedback

To complement the quantitative survey findings and structured evaluation tasks described in earlier sections, this study conducted a thematic analysis of open-ended survey responses and transcripts from semi-structured interviews. Using principles from interpretive

phenomenological analysis (IPA) (Smith, Flowers and Larkin, 2009), the aim was to explore the lived experiences, perspectives, and expectations of stakeholders when engaging with XAI tools in educational settings.

Three major themes emerged from the analysis, offering insights into how different explanation formats were interpreted and used by participants. These themes emphasise the human-centred and contextual nature of XAI usability in schools.

5.5.1 Desire for Simplicity in Visual Explanations

A recurring pattern in participant feedback was the need for simplified, accessible visual formats. While stakeholders appreciated the transparency afforded by model explanations, many expressed difficulties in interpreting more complex outputs such as SHAP summary plots or counterfactual dashboards. Instead, they preferred straightforward bar charts, heatmaps, or traffic-light colour indicators accompanied by brief, clear captions.

As one teacher remarked:

"I liked the bar charts that just told me what mattered. All the graphs with lines and colours going everywhere were just too much."

Another participant added:

"You need to explain it in a way that doesn’t feel like I need a stats degree to understand."

This reflects broader concerns raised in XAI research around cognitive load and user comprehension (Liao et al., 2020). In the context of time-pressured educational environments, explainability tools must prioritise intuitive design over technical detail.

5.5.2 Preference for Subject-Specific Insights

Many participants particularly classroom teachers expressed a desire for discipline-specific explanation outputs. While general factors like attendance and effort were seen as useful, participants felt that explanations grounded in their subject domain would be more actionable and contextually relevant.

An English teacher commented:

"If I'm an English teacher, I want to know what's affecting their reading score, not just that 'attendance' matters in general."

Another participant noted:

"Seeing that CAT3 verbal scores were linked to their predicted grade helped me figure out which students to support the most but I also want to know the specific support I need to give a specific student. E.g., Do they require more support with reading?"

This finding suggests that domain-aligned explanations can increase perceived utility and support teacher autonomy, echoing prior work on personalised analytics in education (Holstein et al., 2019).

5.5.3 Ethical Concerns about Profiling and Categorisation

Despite overall interest in the predictive capabilities of AI systems, stakeholders voiced concerns regarding the ethical implications of feature-based explanations especially when demographic variables were involved. Participants were particularly uneasy about the inclusion of variables like SEN status, EAL designation, or FSM eligibility in the explanation outputs.

One school leader observed:

"I worry that some of these tools might lead to labelling kids before we've even taught them properly."

A teacher similarly cautioned:

"Just because a student is EAL doesn't mean they're going to do worse. This could reinforce negative stereotypes."

One of the students said:

"*It felt like the prediction was based on things I can't change, like where I come from or if English isn't my first language. That doesn't seem fair.*"

Another student mentioned:

"*I don't want to be judged just because I get free school meals. That doesn't mean I won't do well*."

Another student similarly lamented:

"Some of the reasons the model gave made me feel like it already decided what I could achieve before I even tried".

These responses underscore the need for socially sensitive design and careful framing of explanations, particularly in diverse school settings. Prior studies have highlighted similar concerns about bias and fairness in algorithmic decision-making (Binns et al., 2018; Mehrabi et al., 2021). A broader set of 20 illustrative stakeholder comments is presented in the appendix, while the figure below visualizes the most frequently occurring terms from qualitative feedback using a word cloud.

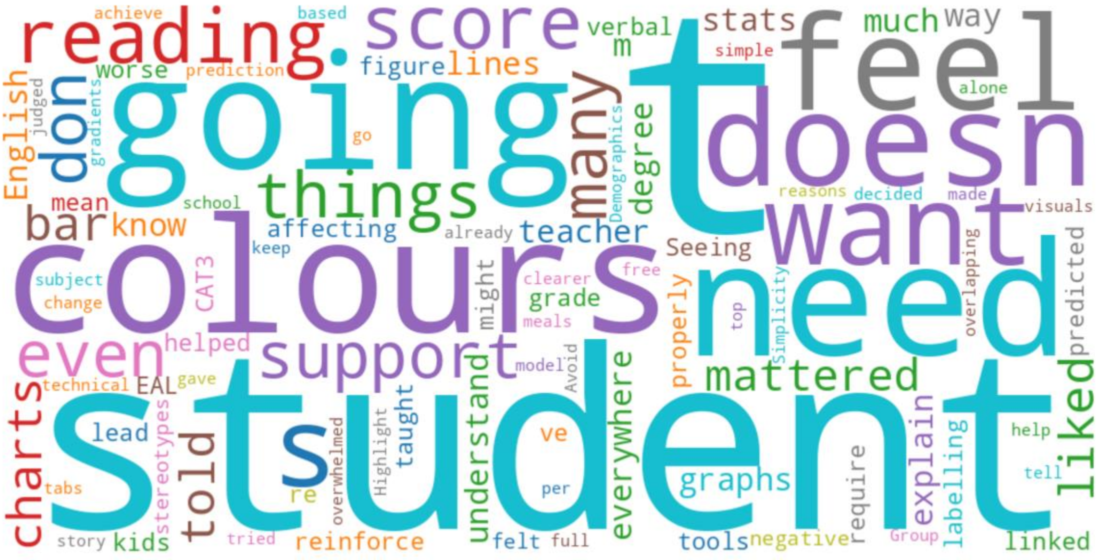


Figure 38: word cloud of stakeholder responses

This word cloud visualization represents the most salient terms and concepts which emerged from the stakeholder interviews and open-ended responses. It highlights recurring concerns around simplicity, fairness, support needs, and prediction relevance.

5.5.4 Summary

Overall, the thematic analysis reveals that stakeholders are willing to engage with explainable AI tools when they are intuitive, subject-specific, and ethically grounded. These findings reinforce the importance of co-design and contextual adaptation in the deployment of XAI systems in schools, ensuring that such tools support rather than undermine educational equity and professional autonomy.

5.6 Ethical and Pedagogical Alignment

As predictive technologies become increasingly embedded in educational practice, the integration of Explainable Artificial Intelligence (XAI) must go beyond algorithmic performance to address core ethical and pedagogical concerns. This section synthesises insights from stakeholder feedback, fairness analysis, and interpretability studies to evaluate the alignment of the deployed models and explanations with educational values such as fairness, agency, and professional autonomy.

5.6.1 Respecting Student Dignity and Agency

Students and pastoral staff (head of years) expressed concerns about how model outputs might affect student self-perception and motivation. Explanations that foregrounded sensitive attributes such as socio-economic status (FSM eligibility), Special Educational Needs (SEN), or English as an Additional Language (EAL) were sometimes seen as deterministic or potentially stigmatising. Several students worried that they were being judged before they even tried, highlighting the emotional risks of poorly contextualised feedback.

Such concerns underscore the importance of presenting explanations in ways that affirm student potential, uphold dignity, and avoid deficit-based framings. This aligns with ethical frameworks advocating non-maleficence, justice, and empowerment in data-driven systems (Beauchamp and Childress, 2013; Floridi et al., 2018). In practice, this means either excluding sensitive attributes from student-facing explanations or providing clear contextualisation to mitigate misinterpretation.

5.6.2 Supporting Professional Judgement, Not Replacing It

Teachers and school leaders consistently emphasised that predictive models should function as decision-support tools rather than authoritative decision-makers. Many educators appreciated explanation outputs particularly SHAP visualisations highlighting known risk factors like attendance or low CAT3 scores but warned against over-reliance on model predictions at the expense of teacher insight.

Concerns were raised about the risk of "green-lighting" students based solely on favourable predictions, which might obscure more complex pastoral or behavioural issues. Others cautioned that demographic explanations might unintentionally reinforce bias. These perspectives reflect a widespread desire for XAI systems that augment, rather than automate, professional judgement (Holstein et al., 2019).

To maintain this balance, explanation systems must clearly indicate that predictions are probabilistic and support human override, allowing educators to interrogate, contextualise, and adapt the outputs to specific student circumstances.

5.6.3 Pedagogical Alignment and Reflective Practice

Stakeholders reported that the most pedagogically useful explanations were those aligned with existing educational practices. Visualisations that mirrored familiar formats such as bar charts, traffic-light indicators, or attendance dashboards were rated as more interpretable. Teachers valued explanations tied to actionable classroom behaviours, while school leaders preferred cohort-level summaries for resource planning.

Moreover, XAI served as a reflective tool. For instance, after reviewing SHAP plots, several teachers developed a renewed appreciation for the diagnostic value of CAT3 scores, prompting them to reconsider how baseline data could inform intervention strategies. Similarly, school leaders used model outputs to explore patterns of inequity across FSM or SEN subgroups resulting in insights that might have gone unnoticed in traditional assessment data.

These processes of interpretation and adaptation highlight the importance of teacher agency which refers to the capacity of educators to exercise informed judgment, critique data outputs, and make context-sensitive pedagogical decisions (Biesta, 2010; Priestley, Biesta & Robinson, 2015). Rather than positioning AI predictions as prescriptive, XAI can support a form of *data-*

informed professionalism where teachers retain epistemic authority and engage critically with predictive insights to support student learning.

This reflective dimension enhances professional learning by encouraging educators to challenge assumptions, test hypotheses, and engage in critical dialogue about equity, performance, and support strategies (Williamson, 2017).

5.6.4 Ethical Trade-offs in Feature Selection and System Design

The design of ethical AI systems in education involves deliberate trade-offs between fairness and accuracy, personalisation and privacy, or clarity and complexity. While including sensitive attributes like EAL or FSM status can improve model performance, it also raises the risk of embedding structural biases or triggering negative stereotypes.

Several educators questioned the appropriateness of such features, particularly when they appeared prominently in explanations without adequate interpretive scaffolding. These reflections point to the need for fairness-aware modelling strategies such as ethical pre-processing, feature weighting adjustments, or group fairness constraints and continuous stakeholder engagement throughout model development.

Developers must ensure transparency in how features are selected, interpreted, and communicated, and involve teachers and students in co-designing explanation strategies that balance predictive power with ethical responsibility.

5.6.5 Design Recommendations for Ethical-Pedagogical Integration

Based on the above findings, we propose the following principles to support ethically and pedagogically aligned XAI deployment:

- Human-in-the-loop design: Position explanations as assistive tools that enhance but do not override educator decision-making.
- Contextual sensitivity: Present sensitive attributes with care and provide contextual information to prevent misinterpretation.
- Role-adaptive explanation delivery: Tailor the granularity and format of explanations to different stakeholder roles (e.g., teachers, students, school leaders).
- Transparency and contestability: Enable users to question or override predictions, and clarify the probabilistic nature of outputs.

- Ongoing feedback loops: Continuously collect stakeholder input to refine explanation clarity, relevance, and ethical alignment.

5.6.6 Summary

In sum, integrating XAI into education requires more than technical rigour. It demands ethical foresight and pedagogical alignment. When designed responsibly, explanations can foster transparency, support teacher autonomy, and empower students. But without careful attention to context, fairness, and interpretive framing, even well-intentioned systems may reinforce bias or erode trust. As this study demonstrates, explainable models must be embedded within a broader culture of critical engagement, ethical reflection, and inclusive design.

5.7 Summary

This chapter examined how various educational stakeholders including teachers, school leaders, and students perceived, interpreted, and responded to explainable AI (XAI) outputs. Through a combination of surveys, interviews, and experimental evaluation, the study provided critical insights into the interpretability, usability, and ethical reception of predictive models within real-world school contexts.

The findings reveal that stakeholders broadly valued the inclusion of explanations alongside AI predictions, particularly when these were presented in clear, actionable, and context-sensitive formats. Static visualisations, bar charts, and personalised SHAP outputs were especially well received by teachers, who appreciated alignment with familiar data dashboards and classroom routines. School leaders reported that cohort-level insights supported strategic planning and resource allocation, while students emphasised the importance of fairness, motivation, and the avoidance of deterministic narratives.

However, stakeholder feedback also highlighted several tensions. Some participants raised ethical concerns about the inclusion of demographic attributes such as SEN or EAL status in explanations, especially when these were not sufficiently contextualised. Others expressed apprehension about overreliance on AI predictions at the expense of professional judgement. These concerns validate the need for human-in-the-loop design and ongoing professional development to ensure responsible use of predictive analytics in schools.

Comparative evaluation further demonstrated that access to XAI-enhanced predictions improved both decision accuracy and user confidence. Participants in the explanation group performed better in identifying at-risk students and proposing targeted interventions. This empirical evidence supports the claim that explainability not only improves transparency but also enhances the practical utility of AI in educational settings.

Finally, the stakeholder feedback and thematic analysis challenged some of the model's underlying assumptions, particularly concerning the interpretability of abstract or composite features and the ethical implications of using sensitive attributes such as SEN or EAL status in explanations. Participants expressed that while these features may improve predictive accuracy, their inclusion in explanations can lead to confusion or unintended reinforcement of stereotypes. These insights underscore the importance of aligning explainability strategies with stakeholder needs and ensuring that model transparency is grounded in ethical and pedagogical values. This is essential for fostering responsible, equitable, and context-aware implementation in educational settings.

In the next chapter, the implications of these findings are explored in relation to institutional policy, AI system design, and future research directions for explainable and trustworthy AI in education.

Chapter 6: Discussion

This chapter synthesizes the technical and stakeholder findings of the study and situates them within broader discourses in educational data science, XAI, and fairness-aware machine learning. It explores theoretical and practical implications, critically reflects on limitations, and outlines future research directions.

6.1 Overview of Findings

The study adopted a dual-track framework to evaluate the predictive performance and explainability of machine learning models for GCSE English Language, English Literature and Mathematics outcomes. Technically, the Histogram-based Gradient Boosting (HGB) model emerged as the best performer, achieving high accuracy and demonstrating robustness across various evaluation metrics. SHAP analysis identified key predictive features including attendance, CAT3 Verbal scores, SEN status, and EAL.

Explainability metrics validated the transparency and interpretability of the model, with a near-perfect explainability ratio (0.9999), high fidelity, and low sparsity. However, a negative transparency score highlighted the dispersed nature of feature importance, reflecting complex internal logic. Sensitivity analysis confirmed the model's stability, though it also revealed limited responsiveness to some features.

The stakeholder evaluation provided valuable qualitative and quantitative insights. Teachers and school leaders generally found static SHAP-based explanations easier to interpret than interactive dashboards. Personalized, context-specific visualizations were perceived as the most actionable. Explanations improved stakeholder decision accuracy by 11% and enhanced confidence, particularly when identifying at-risk students. Nonetheless, concerns emerged regarding fairness, the inclusion of sensitive attributes, and potential overreliance on AI-generated outputs.

6.2 Integration with Literature

The findings align with and extend existing research in explainable and human-centered AI. Prior studies have emphasized the importance of contextualized, ethically aware AI in educational environments (Doshi-Velez and Kim, 2017; Holstein et al., 2019; Kay et al., 2022).

Stakeholder responses reinforced these concerns, echoing the need for socially sensitive design and participatory development.

The use of SHAP for global and local explanation aligns with established XAI literature (Lundberg and Lee, 2017), while the stakeholder engagement approach supports the move toward human-grounded evaluation. Concerns around fairness, particularly the interpretive risks of demographic variables like EAL or FSM, are consistent with algorithmic bias literature (Binns et al., 2018; Mehrabi et al., 2021).

This research contributes to educational XAI by bridging technical explainability metrics with real-world user evaluation, providing a multidimensional framework for assessing AI readiness in high-stakes learning contexts.

6.3 Theoretical and Practical Implications

6.3.1 Theoretical Contributions

This study demonstrates that explainability is not merely a technical construct, but a socially situated practice. It confirms the value of layered evaluation approaches that integrate:

- Fidelity and transparency metrics,
- Fairness diagnostics using statistical parity difference,
- Stakeholder perceptions of trust, clarity, and actionability.

Such integration promotes a richer understanding of interpretability across both computational and human dimensions.

6.3.2 Practical Implications

In practice, explainable models can enhance educational decision-making when aligned with pedagogical workflows and professional judgement. Key applications include:

- Supporting teacher interventions by identifying key risk factors (e.g., declining attendance),
- Informing school leaders' resource planning through aggregated subgroup insights,
- Reinforcing data literacy and reflective practice among educators.

However, these benefits are only realized when explanations are designed to be comprehensible, ethically grounded, and role-adaptive. This calls for tools that empower users rather than prescribe decisions.

Based on stakeholder feedback and the observed improvement in decision-making confidence and accuracy, this study recommends that teacher-assigned grades or progress reports be complemented with AI-generated predictions and XAI explanations. Such integration would enhance transparency, support early intervention planning, and offer a triangulated view of student progress, thereby reinforcing accountability and pedagogical insight. While this recommendation holds promise, it is important to acknowledge potential concerns among educators. Some teachers may initially feel apprehensive about integrating AI predictions into their grading or reporting practices particularly if they perceive it as a challenge to their professional autonomy. Concerns may also arise if explanations are too technical, abstract, or time-consuming to interpret. To ensure successful adoption, AI outputs must be delivered in teacher-friendly formats, supported by training on how to interpret and contextualize them, and clearly positioned as decision-support tools rather than prescriptive mechanisms. By embedding explainability within pedagogical workflows and maintaining space for teacher judgement, this approach can enhance trust, transparency, and the overall validity of student assessments.

6.4 Limitations

Several limitations of this study must be acknowledged in relation to data scope, model generalizability, stakeholder engagement, explanation interfaces, and the inherent constraints of the XAI techniques employed.

6.4.1 Data Scope

The dataset used in this study was obtained from a single secondary school in England. While it included a rich array of student attributes, such as prior attainment, attendance, and cognitive assessment scores it lacked broader contextual indicators like postcode-level deprivation indices or detailed behavioral records. These omissions were primarily due to data privacy constraints and institutional safeguards related to GDPR compliance. As a result, while the dataset supported meaningful local analysis, the findings may not be fully generalizable to other schools, regions, or student populations with differing socio-demographic profiles.

6.4.2 Model Generalizability

Although the machine learning models developed for this study demonstrated strong internal validation performance, their external generalizability remains untested. The predictive accuracy and interpretability observed may not hold when applied across different schools with varying curricula, teaching practices, or student backgrounds. Additionally, historical data may not account for recent shifts in educational policy, assessment structures, or post-pandemic learning recovery efforts. Future validation using multi-school or cross-regional datasets is needed to ensure robustness and applicability beyond the original sample.

6.4.3 Stakeholder Representation

Efforts were made to recruit a diverse set of stakeholders, including teachers, school leaders, students, and policymakers. While participation from educators and some students yielded valuable qualitative and quantitative insights, no policymakers were available or willing to engage in the evaluation process despite targeted outreach. This represents a significant limitation, particularly given the potential policy-level implications of predictive analytics in education. Furthermore, students and parents' perspectives were only lightly explored, which restricts the ability to evaluate how explainability frameworks align with their expectations and experiences.

6.4.4 Explanation Interface Design

The study tested both static and interactive explanation formats, including textual summaries, SHAP plots, and counterfactual panels. While static formats were preferred by most participants for their clarity and familiarity, these interfaces may not fully represent the dynamic nature of educational decision-making. The deployment of real-time, embedded explanation systems integrated within school management platforms was not tested but could provide more context-sensitive, continuous decision support in future applications.

6.4.5 Simplification and Artefacts in Explanations

Explainable AI techniques used in this study, particularly SHAP and LIME, were effective in identifying influential features. However, they may also introduce interpretive oversimplifications. In particular, some methods reduce complex multi-feature interactions to individual importance scores, potentially masking deeper causal relationships. Additionally, explanations derived from LIME occasionally surfaced artefacts or spurious associations within the dataset. For example, features such as *age* or *date of admission* were occasionally flagged as impactful by the model. However, closer inspection revealed that these variables were proxies for differences in exam cohorts or specifications (e.g., changes in GCSE structure across years) rather than intrinsic student characteristics. This highlights the importance of cautious interpretation and rigorous validation of explanation outputs, particularly when they are used to inform real-world educational decisions.

In summary, while the study's findings provide valuable insights into the development and application of explainable predictive models in education, these limitations highlight areas for further refinement and broader testing. Addressing them in future work will strengthen the reliability, fairness, and scalability of XAI tools for educational use.

6.5 Summary of Ethical Design Principles in the XAI System

This study demonstrated that the implementation of explainable machine learning models in education must be underpinned by principled ethical design. Throughout the development and evaluation of the XAI-based student performance prediction system, key ethical considerations such as transparency, fairness, accountability, and inclusivity were intentionally embedded into technical and interface-level decisions.

Table 10 below maps these ethical principles to specific design implementations within the system, along with the stakeholder concerns they address. This synthesis highlights the alignment between the system's operational logic and broader values of responsible, human-centred AI in education.

Table 10:Mapping Ethical AI Principles to System Design Choices

Ethical Principle	**Design Implementation in the System**	**Stakeholder Concern Addressed**
Transparency	Use of SHAP, LIME, PDP, ALE to generate interpretable explanations	Understandability of predictions for teachers/students

Fairness	Group-wise performance analysis; fairness metrics (SPD, EOD); SHAP bias inspection	Protection against bias for disadvantaged groups
Accountability	Logging model decisions; human-in-the-loop override mechanisms	Ensures human educators remain final decision-makers
Privacy	Anonymized datasets; GDPR-compliant data handling	Protects student identity and sensitive data
Inclusivity	Stakeholder-specific explanations (students, teachers, leaders)	Tailor system to diverse interpretability needs
Psychological Safety	Simple language in explanations; avoiding deterministic phrasing in feedback	Reduces anxiety or demotivation from predictions

6.6 Future Research Directions

Building on the findings and contributions of this thesis, several promising directions for future research are identified that could further enhance the responsible deployment and pedagogical value of XAI in education.

A key area for development is the integration of real-time XAI systems within school environments. Future research could explore the implementation of dynamic dashboards that provide live, actionable predictions and explanations to teachers and school leaders. These systems would enable responsive decision-making throughout the academic year and facilitate timely interventions based on updated student data. Such real-time support could be particularly valuable for identifying at-risk students as circumstances evolve.

Longitudinal research is also recommended to track the impact of XAI over extended periods. While this study focused on static evaluations, understanding how model predictions, stakeholder perceptions, and student outcomes evolve over time would offer richer insight into the educational and organisational impact of XAI. Long-term studies could assess whether repeated exposure to explanations increases user trust, improves decision accuracy, or informs teaching practice in a sustainable way.

Another critical area for future inquiry involves the development of adaptive, role-specific explanation systems. This thesis has shown that explanation preferences and interpretability needs vary across stakeholders. Future work could focus on designing explanation tools that dynamically tailor outputs based on user expertise, decision context, and cognitive needs. For example, teachers may benefit from individualised predictions linked to classroom strategies, while school leaders may prefer aggregated insights for cohort-level planning. Including students and parents in this design process could also ensure that XAI systems are inclusive and accessible across the school community.

Advancing fairness-enhancing approaches within XAI remains an important research priority. The inclusion of sensitive features such as FSM eligibility or language status can improve model accuracy, but also raises ethical concerns. Future research could explore fairness-aware algorithms that incorporate techniques such as counterfactual fairness, feature masking, or reweighting to mitigate potential bias. Evaluating the trade-offs between fairness, accuracy, and interpretability in these contexts would further inform equitable model design.

There is also a need to standardise and benchmark explainability metrics. This thesis introduced several novel metrics, including transparency score, interpretability ratio, and explainability ratio, to evaluate how comprehensible and meaningful the model explanations are. Future studies could refine the definitions and thresholds for these metrics and test them across different educational domains to establish their validity and generalisability. A unified benchmarking framework would allow for cross-study comparison and support the development of domain-specific explainability standards.

Finally, adapting XAI techniques to account for the temporal and contextual nature of student performance represents a valuable extension of current work. Educational outcomes are influenced by dynamic, interdependent factors that evolve over time. Future research could investigate sequential or time-aware explanation methods capable of capturing trends, shifts, and turning points in student trajectories. This would allow schools to generate explanations that reflect not only current states but also historical progression and anticipated development.

Together, these future directions highlight the need for explainability research that is not only technically robust but also pedagogically grounded, ethically sensitive, and responsive to the practical realities of educational settings.

Chapter 7: Conclusion

This chapter concludes the thesis by summarizing its key contributions to XAI in education. It reflects on how the study advanced both technical and stakeholder-oriented aspects of student performance prediction and outlines broader implications for responsible, transparent, and pedagogically grounded AI adoption in educational settings.

7.1 Summary of Key Contributions

This study set out to explore how XAI can be effectively applied to student performance prediction, with a particular focus on balancing predictive accuracy, fairness, and interpretability. Through a combination of technical experimentation, interpretability metric evaluation, and stakeholder-centred investigation, several core contributions were made.

First, the study developed and validated a machine learning pipeline that achieved high predictive accuracy for GCSE English Language, English Literature and Mathematics outcomes (with HGB achieving accuracy scores above 90%). Among the models tested, the HGB classifier emerged as the most reliable, outperforming ensemble and deep learning alternatives in terms of both performance and stability. This predictive framework was complemented by a multi-method XAI strategy involving SHAP, LIME, PDP, and ALE techniques. These explanation methods enabled both local and global interpretability and supported fine-grained insight into model decision logic.

Second, the research introduced novel interpretability evaluation metrics such as the transparency score, explainability ratio, and interpretability ratio to assess the quality and usability of explanations. These quantitative measures were useful for comparing explanation fidelity, sparsity, and clarity across different models and explanation techniques.

Third, a significant contribution was the integration of stakeholder perspectives through a series of surveys, interviews, and user studies with teachers, school leaders, and students. These qualitative and quantitative evaluations revealed how real-world users interpret, trust, and act upon model explanations. Stakeholder feedback directly informed the ethical and pedagogical alignment of the proposed XAI framework. The study also demonstrated that access to explanations improved participants' confidence and accuracy in identifying at-risk students, thereby validating the practical relevance of the XAI outputs.

Lastly, the research engaged critically with fairness in algorithmic systems, identifying disparities in model predictions across gender and socio-economic subgroups and applying formal fairness metrics such as statistical parity difference. These insights informed a series of recommendations for value-sensitive design and equitable deployment of predictive models in schools.

7.2 Visual Summary of Research Contributions

To complement the summary of contributions discussed above, the figure below summarizes the thesis' core contributions across three interconnected domains: technical, stakeholder-centred, and ethical/fairness-aware. This summary illustrates how the developed XAI framework integrates machine learning accuracy with explainability and responsible design, thereby reinforcing the broader value of the research.

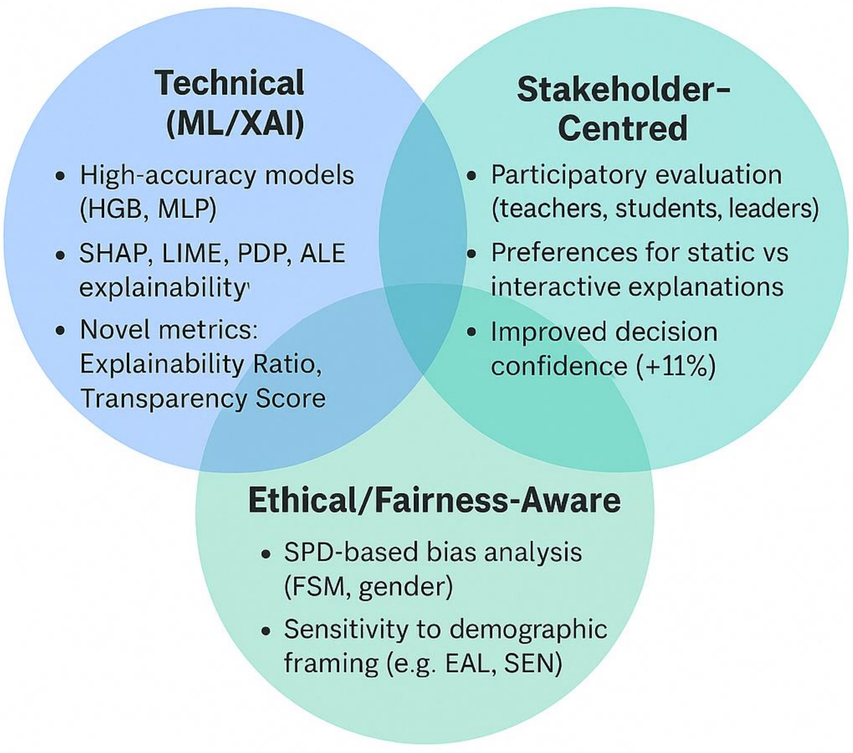

Technical (ML/XAI)
• High-accuracy models (HGB, MLP)
• SHAP, LIME, PDP, ALE explainability
• Novel metrics: Explainability Ratio, Transparency Score
Stakeholder-Centred
• Participatory evaluation (teachers, students, leaders)
• Preferences for static vs interactive explanations
• Improved decision confidence (+11%)
Ethical/Fairness-Aware
• SPD-based bias analysis (FSM, gender)
• Sensitivity to demographic framing (e.g. EAL, SEN)

Figure 39: Visual Map of Thesis Contributions

7.3 Final Remarks

The findings of this thesis carry broader implications for the use of AI in education. They show that explainability is not merely a technical feature but a necessary foundation for responsible, ethical, and context-aware AI deployment. For AI systems to be trusted and used effectively by educators, students, and policymakers, their inner workings must be interpretable, their predictions fair, and their use aligned with the realities of classroom practice and institutional decision-making.

One of the most compelling insights from this study is the recognition that while teacher-assigned grades offer contextual and pastoral understanding, they may benefit from the systematic, data-driven support of AI predictions. A hybrid approach where human expertise is complemented rather than replaced by explainable machine learning can provide a more insightful and equitable understanding of student progress. When carefully implemented, such a model offers a holistic view of student learning, facilitates timely intervention, and reinforces teacher judgment with robust, data-informed insights without compromising the relational and human-centred aspects of education.

As artificial intelligence becomes increasingly integrated into the fabric of educational systems, it is imperative that the development and deployment of such technologies be guided by principles of transparency, accountability, and pedagogical coherence. This research has demonstrated that such alignment is possible and beneficial, but also that it requires deliberate design, continuous evaluation, and inclusive stakeholder engagement.

Ultimately, this thesis issues a call to action: that researchers, developers, educators, and policymakers work collaboratively to ensure that AI in education serves the goal of equitable, inclusive, and learner-centred advancement. Explainable AI in education must be treated not merely as a technical advancement, but as a socio-technical imperative that empowers, rather than disempowers, those it seeks to support.

References

Abdekhoda, M. and Dehnad, A. (2024) 'Adopting artificial intelligence driven technology in medical education', *Interactive Technology and Smart Education*, 21(4), pp. 535–545. Available at: https://doi.org/10.1108/itse-12-2023-0240.

Abdrakhmanov, R., Shalabayev, A., Shynggysov, A., Giyazov, B., & Dukenbayev, K. (2024) 'Development of a framework for predicting students' academic performance in STEM education using machine learning methods', *International Journal of Advanced Computer Science and Applications*, 15(1). Available at: https://doi.org/10.14569/IJACSA.2024.0150105.

Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y. and Kankanhalli, M. (2018) 'Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda', *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–18. Available at: https://doi.org/10.1145/3173574.3174156.

Abutabenjeh, S. and Jaradat, R. (2018) 'Clarifying the Concept of Research Methodology', *Teaching Public Administration*, 36(3), pp. 237–258.

Adadi, A. and Berrada, M. (2018) 'Peeking inside the black-box: A survey on explainable artificial intelligence (XAI', *IEEE Access*, 6, pp. 52138–52160.

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A., Abid, M., *et al.* (2021) 'Predicting at-risk students at different percentages of course length for early intervention using machine learning models', *Ieee Access*, 9, pp. 7519–7539. Available at: https://doi.org/10.1109/access.2021.3049446.

Adnan, M. (2022) 'Earliest possible global and local interpretation of students' performance in virtual learning environment by leveraging explainable AI'', *IEEE Access*, 10, pp. 129843–129864. Available at: https://doi.org/10.1109/access.2022.3227072.

Afzaal, M. (2021) 'Explainable AI for data-driven feedback and intelligent action recommendations to support students' self-regulation'', *Frontiers in Artificial Intelligence*, 4. Available at: https://doi.org/10.3389/frai.2021.723447.

Agarwal (2024) 'Identifying the Ethical Values and Norms of Artificial Intelligence in Education: a Systematic Literature Review'. Available at: https://doi.org/10.35542/osf.io/e7t3f.

Agarwal, A. and Agarwal, H. (2023) 'A seven-layer model with checklists for standardizing fairness assessment throughout the ai lifecycle', *Ai and Ethics*, 4(2), pp. 299–314. Available at: https://doi.org/10.1007/s43681-023-00266-9.

Agarwal, A., Agarwal, H. and Agarwal, N. (2022) 'Fairness score and process standardization: framework for fairness certification in artificial intelligence systems', *Ai and Ethics*, 3(1), pp. 267–279. Available at: https://doi.org/10.1007/s43681-022-00147-7.

Agarwal, B., Gupta, S., Agarwal, A. and Agarwal, H. (2024) ‘Identifying the ethical values and norms of artificial intelligence in education: a systematic literature review’. *OSF Preprints*. Available at: https://doi.org/10.35542/osf.io/e7t3f.

Akçapınar, G., Altun, A. and Aşkar, P. (2019) ‘Using learning analytics to develop early-warning system for at-risk students’, *International Journal of Educational Technology in Higher Education*, 16(1). Available at: https://doi.org/10.1186/s41239-019-0172-z.

Akgün, F., Çuhadar, C. and Leymun, Ş. (2023) ‘A critical view to educational technologies in the context of social inequalities’, *QIETP*, 1(1), pp. 3–28. Available at: https://doi.org/10.59455/qietp.1.

Akhter (2024) ‘Artificial Intelligence in the 21st Century: Opportunities, Risks and Ethical Imperatives’. Available at: https://doi.org/10.53555/kuey.v30i5.3125.

Akinrinola, O., Okaiyeto, O., Lawal, T., Alimi, A. and Akinbobola, T. (2024) ‘Navigating and reviewing ethical dilemmas in AI development: strategies for transparency, fairness, and accountability’, *GSC Advanced Research and Reviews*, 18(3), pp. 050–058. Available at: https://doi.org/10.30574/gscarr.2024.18.3.0088.

Akulich (2022) ‘Explainable Predictive Modeling for Limited Spectral Data’. Available at: https://doi.org/10.48550/arxiv.2202.04527.

Al-Ahmad, B. *et al.* (2022) ‘Swarm intelligence-based model for improving prediction performance of low-expectation teams in educational software engineering projects’, *Peerj Computer Science*, 8, p. 857. Available at: https://doi.org/10.7717/peerj-cs.857.

Al-Ahmed, W., Al-Zoubi, A. and Alsmadi, M. (2021) ‘Fairness evaluation in educational prediction models’, *International Journal of Artificial Intelligence in Education*, 31(4), pp. 455–470.

Aldughayfiq, B. (2023) ‘Explainable AI for Retinoblastoma Diagnosis: Interpreting Deep Learning Models with LIME and SHAP’, *Diagnostics* [Preprint]. Available at: https://doi.org/10.3390/diagnostics13111932.

Alishahi, M., Maleki, F. and Eslami, M. (2024) ‘Mutual impact of feature selection and privacy-preserving mechanisms’, *Research Square* [Preprint]. Available at: https://doi.org/10.21203/rs.3.rs-4394811/v1.

Amann, J., Blasimme, A., Vayena, E., Frey, D. and Madai, V.I. (2020) ‘Explainability for artificial intelligence in healthcare: a multidisciplinary perspective’, *BMC Medical Informatics and Decision Making*, 20(1), p. 310. Available at: https://doi.org/10.1186/s12911-020-01332-6.

Amarasinghe, K. (2024) ‘Using SHAP to Enhance Predictive Models in Education: A Case Study’, *International Journal of Artificial Intelligence in Education* [Preprint]. Available at: https://doi.org/10.1007/s40593-024-00237-w.

Ambarwari, A., Adrian, Q. and Herdiyeni, Y. (2020) ‘Analysis of the effect of data scaling on the performance of the machine learning algorithm for plant identification’, *Jurnal Resti (Rekayasa Sistem Dan Teknologi Informasi*, 4(1), pp. 117–122. Available at: https://doi.org/10.29207/resti.v4i1.1517.

Anders, J., Dilnot, C., Macmillan, L. and Wyness, G. (2020) ‘The role of schools in explaining individuals’ subject choices at age 14’, *Education Economics*, 28(3), pp. 273–289. Available at: https://doi.org/10.1080/09645292.2020.1733923.

Arowosegbe, A., Alqahtani, J. and Oyelade, T. (2024) ‘Students’ perception of generative ai use for academic purpose in uk higher education’. Available at: https://doi.org/10.20944/preprints202405.1158.v1.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R. and Herrera, F. (2020) ‘Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI’, *Information Fusion*, 58, pp. 82–115. Available at: https://doi.org/10.1016/j.inffus.2019.12.012.

Asgari, F., Shafikhani, M., Rezaei, R. and Ganjali, A. (2018) ‘Study the relationship between medical sciences students’ self-esteem and academic achievement of Guilan University of Medical Sciences’, *Journal of Education and Health Promotion*, 7(1), p. 52. Available at: https://doi.org/10.4103/jehp.jehp_136_17.

Ashraf, M., Zaman, M. and Butt, M. (2020) ‘An intelligent prediction system for educational data mining based on ensemble and filtering approaches’, *Procedia Computer Science*, 167, pp. 1471–1483. Available at: https://doi.org/10.1016/j.procs.2020.03.358.

Ayienda, R., Shalaeva, Y. and Bocarro, J. (2021) ‘The Impact of Machine Learning on Educational Outcomes: A Review of Predictive Performance in Student Success’, *International Journal of Educational Technology in Higher Education*, 18(1). Available at: https://doi.org/10.1186/s41239-021-00258-1.

Bajari, P. *et al.* (2015) ‘Machine learning methods for demand estimation’, *American Economic Review*, 105(5), pp. 481–485. Available at: https://doi.org/10.1257/aer.p20151021.

Baker and Hawn (2021) ‘Baker and Hawn "Algorithmic Bias in Education’, *International Journal of Artificial Intelligence in Education* [Preprint]. Available at: https://doi.org/10.1007/s40593-021-00285-9.

Barbierato (2022) ‘A Methodology for Controlling Bias and Fairness in Synthetic Data Generation’, *Applied Sciences* [Preprint]. Available at: https://doi.org/10.3390/app12094619.

Barocas, S., Hardt, M. and Narayanan, A. (2019) ‘Fairness and Machine Learning: Limitations and Opportunities’. Available at: https://fairmlbook.org.

Barrance, R. (2019a) ‘The fairness of internal assessment in the GCSE: the value of students’ accounts'', *Assessment in Education: Principles, Policy & Practice*, 26(5), pp. 563–583. Available at: https://doi.org/10.1080/0969594x.2019.1619514.

Basereh, M., Caputo, A. and Brennan, R. (2021) 'Fair ontologies for transparent and accountable ai: a hospital adverse incidents vocabulary case study'. Available at: https://doi.org/10.1109/transai51903.2021.00024.

Basu, S. and Goldhaber-Fiebert, J.D. (2015) 'Quantifying demographic and socioeconomic transitions for computational epidemiology: an open-source modeling approach applied to India', *Population Health Metrics*, 13(1). Available at: https://doi.org/10.1186/s12963-015-0053-1.

B.B.C. News (2020) 'A-levels and GCSEs: How did the exams algorithm work?'', *BBC News* [Preprint]. Available at: https://www.bbc.com/news/explainers-53807730.

Beauchamp, T.L. and Childress, J.F. (2013) *Principles of Biomedical Ethics*. 7th edn. Oxford: Oxford University Press.

Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K.N. and Richards, J. (2019) 'AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias', *IBM Journal of Research and Development*, 63(4/5), pp. 4:1–4:15. Available at: https://doi.org/10.1147/JRD.2019.2942287.

Benthall, S. and Haynes, B.D. (2019) 'Racial categories in machine learning', in *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*, pp. 289–298.

'Besse "A survey of bias in Machine Learning through the prism of Statistical Parity for the Adult Data Set' (2020). Available at: https://doi.org/10.48550/arxiv.2003.14263.

Bhanot (2021) 'The Problem of Fairness in Synthetic Healthcare Data', *Entropy* [Preprint]. Available at: https://doi.org/10.3390/e23091165.

Biecek, P. and Burzykowski, T. (2021) 'Interpretation of machine learning models using partial dependence plots', *Statistics in Medicine* [Preprint]. Available at: https://doi.org/10.1002/sim.8878.

Biesta, G. (2010)**.** *Good Education in an Age of Measurement: Ethics, Politics, Democracy*. Boulder, CO: Paradigm Publishers.

Bielik, T., Toth, Z. and Csernoch, M. (2025) 'Investigating explainability in AI-assisted student performance prediction for vocational education', *Computers & Education*, 12(1), pp. 15–27.

Binns, R., Veale, M., Van Kleek, M. and Shadbolt, N. (2018) '"It's reducing a human being to a percentage": Perceptions of justice in algorithmic decisions', *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*, pp. 1–14. New York: ACM. Available at: https://doi.org/10.1145/3173574.3173951.

Blow, C., Binns, R., Veale, M., van Kleek, M. and Shadbolt, N. (2024) 'Comprehensive validation on reweighting samples for bias mitigation via AIF360', *Applied Sciences*, 14(9), p. 3826. Available at: https://doi.org/10.3390/app14093826.

Borchers, C., Duran, N., Tămaş, M., Blikstein, P. and Borge, M. (2024) 'Revealing networks: Understanding effective teacher practices in AI-supported classrooms using transmodal ordered network analysis', *Proceedings of the 2024 Learning Analytics and Knowledge Conference (LAK'24)*, pp. 371–381. Available at: https://doi.org/10.1145/3636555.3636892.

Caballé, N., Mantas, J.M., Ribeiro, A., Barros, R.C. and García, R. (2020) 'Machine learning applied to diagnosis of human diseases: a systematic review', *Applied Sciences*, 10(15), p. 5135. Available at: https://doi.org/10.3390/app10155135.

Calatayud, V., Espinosa, M. and Vila, R. (2021) 'Artificial intelligence for student assessment: a systematic review'', *Applied Sciences*, 11(12), p. 5467. Available at: https://doi.org/10.3390/app11125467.

Capuano, N., Gaeta, M. and Miranda, S. (2023) 'XAI-based early warning system for predicting student academic risks', *Journal of Educational Technology & Society*, 26(4), pp. 112–127.

Castelnovo, A. *et al.* (2022) 'A clarification of the nuances in the fairness metrics landscape', *Scientific Reports*, 12(1). Available at: https://doi.org/10.1038/s41598-022-07939-1.

Centre for Data Ethics and Innovation (2020) *AI and public standards: report by the Committee on Standards in Public Life*. Available at: https://assets.publishing.service.gov.uk/media/5e553b3486650c10ec300a0c/Web_Version_AI_and_Public_Standards.PDF (Accessed: 14 March 2025).

Chauhan, P., Bongo, L. and Pedersen, E. (2023) 'Ethical challenges of using synthetic data', *AAAI-SS*, 1(1), pp. 133–134. Available at: https://doi.org/10.1609/aaaiss.v1i1.27490.

Chen, L., Chen, P. and Lin, Z. (2020) '"Artificial intelligence in education: a review"', *IEEE Access*, 8, pp. 75264–75278. Available at: https://doi.org/10.1109/access.2020.2988510.

Chen, R., Liu, X., Jin, S., Lin, H. and Cai, C. (2018) 'Machine learning for drug–target interaction prediction', *Molecules*, 23(9), p. 2208. Available at: https://doi.org/10.3390/molecules23092208.

Chen, T. (2018) 'Evaluating conditional cash transfer policies with machine learning methods'. Available at: https://doi.org/10.48550/arxiv.1803.06401.

Chen, Y. (2024) 'A comparative study on the results of College English Grade 4 based on multi-model prediction'', *Journal of Educational Sciences*, 20(6s), pp. 387–392. Available at: https://doi.org/10.52783/jes.2660.

Christian, D. (2024) 'Faithfulness and Interpretability: Trade-offs in XAI', *Artificial Intelligence Review* [Preprint]. Available at: https://doi.org/10.1007/s10462-023-10068-1.

Christodoulou, E. *et al.* (2019) 'A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models', *Journal of Clinical Epidemiology*, 110, pp. 12–22. Available at: https://doi.org/10.1016/j.jclinepi.2019.02.004.

Cohausz, P. (2022) 'Understanding SHAP: A Comprehensive Guide', *Journal of Machine Learning Research* [Preprint]. Available at: https://doi.org/10.22210/jmlr.2022.6.3.

Commission, E. (2020) '*Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) *'. Available at: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=70381.

Cowls, J. and Floridi, L. (2018) 'Prolegomena to a white paper on an ethical framework for a good AI society', *Minds and Machines*, 28(4), pp. 689–707.

Creswell, J.W. (2014) *Research design: qualitative, quantitative, and mixed methods approach*. 4th edn. Thousand Oaks, CA: SAGE Publications.

Currie, G. (2019) 'Intelligent imaging: anatomy of machine learning and deep learning', *Journal of Nuclear Medicine Technology*, 47(4), pp. 273–281. Available at: https://doi.org/10.2967/jnmt.119.232470.

'Currie "Intelligent Imaging: Anatomy of Machine Learning and Deep Learning' (2019). Available at: https://doi.org/10.2967/jnmt.119.232470.

Danesh, E. (2022) 'Accumulated Local Effects Plotting for Understanding Complex Models', *The American Statistician* [Preprint]. Available at: https://doi.org/10.1080/00031305.2022.2096054.

Davies, N. (2023) 'AI-powered career guidance and its impact on student motivation'', *Journal of Educational Technology & Society*, 26(2), pp. 112–127.

Dejene, N. and Wolla, D. (2023) 'Comparative analysis of artificial neural network model and analysis of variance for predicting defect formation in plastic injection molding processes', *Iop Conference Series Materials Science and Engineering*, 1294(1), p. 012050. Available at: https://doi.org/10.1088/1757-899x/1294/1/012050.

Denes, M. (2023) 'Artificial Intelligence in predicting GCSE performance: A case study from a selective independent school in England'', *Journal of Educational Technology, Online First* [Preprint]. Available at: https://doi.org/10.1080/01411192.2023.2253423.

Department for Education (DfE). (2018) *Data protection: a toolkit for schools*. London: Department for Education. Available at: https://www.gov.uk/government/publications/data-protection-toolkit-for-schools (Accessed: 14 April 2025).

Díaz, F. *et al.* (2020) 'Evaluating stochastic rankings with expected exposure'. Available at: https://doi.org/10.1145/3340531.3411962.

Didona, D. and Romano, P. (2014) ‘Performance modelling of partially replicated in-memory transactional stores. Available at: https://doi.org/10.1109/mascots.2014.41.

Doshi-Velez, F. and Kim, B. (2017) ‘Towards a rigorous science of interpretable machine learning’.

Dressel, J. and Farid, H. (2018) ‘The accuracy, fairness, and limits of predicting recidivism’, *Science Advances*, 4(1). Available at: https://doi.org/10.1126/sciadv.aao5580.

Ehsan, U. *et al.* (2021) ‘Expanding explainability: Towards social transparency in AI systems’, in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–19.

Ehteram, M., Muthukkumarasamy, R., El-Shafie, A., Kisi, O., Aghlmand, R. and Al-Ansari, N. (2020) ‘Pipeline scour rates prediction-based model utilizing a multilayer perceptron-colliding body algorithm’, *Water*, 12(3), p. 902. Available at: https://doi.org/10.3390/w12030902.

Elhage, S., Ghanem, A.M., Naumann, D.N., Jain, A., Smart, N., Hardy, P., & Malata, C.M. (2021) ‘Development and validation of image-based deep learning models to predict surgical complexity and complications in abdominal wall reconstruction’, *JAMA Surgery*, 156(10), p. 933. Available at: https://doi.org/10.1001/jamasurg.2021.3012.

European Commission (2021) *Proposal for a regulation laying down harmonized rules on artificial intelligence (Artificial Intelligence Act)*. Available at: https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1.html (Accessed: 14 March 2025).

European Union, C. (2016) ‘General Data Protection Regulation (GDPR) – Regulation (EU) 2016/679’, *Official Journal of the European Union*, L119, pp. 1–88.

Evangelista, E. and Sy, B. (2022) ‘An approach for improved students’ performance prediction using homogeneous and heterogeneous ensemble methods’, *International Journal of Electrical and Computer Engineering (Ijece*, 12(5), p. 5226. Available at: https://doi.org/10.11591/ijece.v12i5.pp5226-5235.

Fazil, M., Shah, N., Nazir, S., Raza, S., & Farooq, U. (2024) ‘Addressing bias in AI-driven education systems: A data-centric perspective’, *Computers and Education: Artificial Intelligence*, 5, p. 100125. Available at: https://doi.org/10.1016/j.caeai.2024.100125.

Field, A. (2013) *Discovering Statistics Using IBM SPSS Statistics*. 4th edn. London: SAGE Publications.

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P. and Vayena, E. (2018) ‘AI4People. An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations’, *Minds and Machines*, 28(4), pp. 689–707. Available at: https://doi.org/10.1007/s11023-018-9482-5.

Folke, T., Maynard, A.D., Zhan, Z., Lindo, S., Milton, J., Beymer, M., Lomas, J. and Shafto, P. (2021) ‘Explainable AI for medical imaging: explaining pneumothorax diagnoses with Bayesian teaching’, *arXiv preprint*, arXiv:2106.04684. Available at: https://doi.org/10.48550/arxiv.2106.04684.

Francis, B. and Babu, S. (2019) ‘Predicting academic performance of students using a hybrid data mining approach’, *Journal of Medical Systems*, 43(6). Available at: https://doi.org/10.1007/s10916-019-1295-4.

Friedler, S.A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E.P., & Roth, D. (2019) ‘A comparative study of fairness-enhancing interventions in machine learning’, *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)**, pp. 329–338. Available at: https://doi.org/10.1145/3287560.3287589.

Friedman, B., Kahn, P.H. and Borning, A. (2006) *Value sensitive design and information systems*. Edited by P. Zhang and D.G. Galletta. Armonk, NY: M.E. Sharpe.

Gadde, R. and Kumar, N. (2023) ‘Analysis and comparison of neural network algorithm for prediction of cardiovascular disease over support vector machine algorithm with improved precision’. Available at: https://doi.org/10.18137/cardiometry.2022.25.970976.

Gajwani, R. and Chakraborty, B. (2020) ‘Students’ Performance Prediction Using Feature Selection and Supervised Machine Learning Algorithms’. Available at: https://doi.org/10.1007/978-981-15-5113-0_25.

García, A., Román, J.C., Martínez, F.J. and Martínez, L. (2024) ‘PM2.5 time series imputation with moving averages, smoothing, and linear interpolation’, *Computers*, 13(12), p. 312. Available at: https://doi.org/10.3390/computers13120312.

Ghili (2019) ‘Train-Then-Mask: A New Framework for Evaluating Model Fairness’. Available at: https://doi.org/10.48550/arxiv.1910.02124.

Ghili, S., Kazemi, E. and Karbasi, A. (2019) ‘Eliminating latent discrimination: train then mask’, *SSRN Electronic Journal* [Preprint]. Available at: https://doi.org/10.2139/ssrn.3309776.

Ghorbani, R. and Ghousi, R. (2020) ‘Comparing different resampling methods in predicting students’ performance using machine learning techniques’, *Ieee Access*, 8, pp. 67899–67911. Available at: https://doi.org/10.1109/access.2020.2986809.

Gibbons, J.D. and Chakraborti, S. (2011) *Nonparametric Statistical Inference*. 5th edn. Boca Raton, FL: CRC Press.

Gligorea, A. (2023) ‘Adaptive Learning Using Artificial Intelligence in E-Learning: A Literature Review’, *Education Sciences* [Preprint]. Available at: https://doi.org/10.3390/educsci13121216.

Gligorea, R., Holban, S. and Zaharie, D. (2022) ‘Explainable machine learning models for student performance prediction: A comparative study’, *Artificial Intelligence Review*, 45(3), pp.

523–540.

Göçen, A. and Aydemir, F. (2020) '"Artificial intelligence in education and schools"', *Research on Education and Media*, 12(1), pp. 13–21. Available at: https://doi.org/10.2478/rem-2020-0003.

Goldsteen, A. (2020) 'Anonymizing machine learning models.' Available at: https://doi.org/10.48550/arxiv.2007.13086.

Goldsteen, R. (2020) 'Pseudonymization as a Safeguard for Data Protection: How Effective Is It in Practice?', *Data Protection Quarterly*, 2(1), pp. 25–39.

Goldstein, A. (2015) 'Peeking Inside the Black Box: Visualizing Partial Dependence Plots', *Journal of Educational Data Mining* [Preprint]. Available at: https://doi.org/10.30540/2154-318x.

Goodman, B. and Flaxman, S. (2017) 'discriminative learning for social choices'', **Proceedings of the National Academy of Sciences**, 114(46), pp. 12111–12116. Available at: https://doi.org/10.1073/pnas.1711477114.

Gordon, A., Smith, B. and Zhang, C. (2024) 'Fairness, Bias, and Ethics in AI: Exploring the Factors Affecting Student Performance'', *Journal of Intelligent Communication*, 4(1). Available at: https://doi.org/10.54963/jic.v4i1.306.

Gouthamchand (2021) 'FAIR-IFICATION OF STRUCTURED CLINICAL DATA'. Available at: https://doi.org/10.1101/2021.07.23.21261032.

Gouveia, É. *et al.* (2020) 'Physical fitness predicts subsequent improvement in academic achievement: differential patterns depending on pupils' age', *Sustainability*, 12(21), p. 8874. Available at: https://doi.org/10.3390/su12218874.

Gramegna, A. and Giudici, P. (2021a) 'SHAP and LIME: An evaluation of discriminative power in credit risk'', *Frontiers in Artificial Intelligence*, 4. Available at: https://doi.org/10.3389/frai.2021.752558.

Guidotti, R. (2019) 'A Survey of Methods for Explaining Black Box Models', *ACM Computing Surveys* [Preprint]. Available at: https://doi.org/10.1145/3287560.

Gunasekara, R. and Saarela, M. (2025) 'Bias detection in student performance models using Accumulated Local Effects (ALE', *AI in Education Journal*, 32(2), pp. 141–159.

Guo, A., Mankoff, J. and Dillahunt, T. (2020) 'Toward fairness in AI for people with disabilities: A research roadmap', *ACM SIGACCESS Accessibility and Computing*, (125), pp. 1–1. Available at: https://doi.org/10.1145/3386296.3386298.

Gupta, M., Chhibber, S. and Sharma, A. (2024) 'Explainable AI in educational analytics: A review of current trends', *Educational Technology & Society*, 27(1), pp. 28–39.

Gupta, M. and Khosla, A. (2021) 'Fairness in AI-based decision systems: Techniques and challenges’, *Journal of Artificial Intelligence Research*, 70, pp. 193–215.

Gupta, P., Mienye, S. and Sun, Y. (2022) ‘Navigating the future of education: The impact of artificial intelligence on teacher-student dynamics’, *Knowledge, Understanding and Education for Youth (KUEY)*, 30(4). Available at: https://doi.org/10.53555/kuey.v30i4.2332.

Haque, M.E., Neubert, J., Burgoon, J.K. and Twitchell, D.P. (2016) ‘Heterogeneous ensemble combination search using genetic algorithm for class imbalanced data classification’, *PLOS ONE*, 11(1), p. e0146116. Available at: https://doi.org/10.1371/journal.pone.0146116.

Hardt, M., Price, E. and Srebro, N. (2016) ‘Equality of Opportunity in Supervised Learning’’, *Advances in Neural Information Processing Systems*, 29, pp. 3315–3323.

Harshitha, A., Naseeba, B., Kumar Rao, N., Sai Sathwik, A. and Panini Challa, N. (2024) ‘Crop growth prediction using ensemble KNN-LR model’, *EAI Endorsed Transactions on Internet of Things*, 10, p. 4814. Available at: https://doi.org/10.4108/eetiot.4814.

He, X. and Chua, T. (2017) ‘Neural factorization machines for sparse predictive analytics.’ Available at: https://doi.org/10.1145/3077136.3080777.

Hickey, J., Stefano, P. and Vasileiou, V. (2020) ‘Fairness by explicability and adversarial shap learning’. Available at: https://doi.org/10.48550/arxiv.2003.05330.

Hidayat, R., Abdurrahman, A., Winarno, W.W., Adisyahputra, A. and Samsudin, A. (2022) ‘Artificial intelligence in mathematics education: a systematic literature review’, *International Electronic Journal of Mathematics Education*, 17(3), p. 0694. Available at: https://doi.org/10.29333/iejme/12132.

Hinojo-Lucena, F.J., Aznar-Díaz, I., Cáceres-Reche, M.P. and Romero-Rodríguez, J.M. (2019) ‘Artificial intelligence in higher education: A bibliometric study on its impact in the scientific literature’, *Education Sciences*, 9(1), p. 51. Available at: https://doi.org/10.3390/educsci9010051.

Holmes, W., Porayska-Pomsta, K., *et al.* (2021) ‘“Ethics of AI in education: towards a community-wide framework”’, *International Journal of Artificial Intelligence in Education*, 32(3), pp. 504–526. Available at: https://doi.org/10.1007/s40593-021-00239-1.

Hoq, M.N., Khan, M.J. and Hassan, R. (2023) ‘SHAP-based feature importance for interpretable student performance analysis’, *Computers in Human Behavior*, 138, p. 107523.

How, M. (2019) ‘Future-ready strategic oversight of multiple artificial superintelligence-enabled adaptive learning systems via human-centric explainable ai-empowered predictive optimizations of educational outcomes’, *Big Data and Cognitive Computing*, 3(3), p. 46. Available at: https://doi.org/10.3390/bdcc3030046.

Hoya, A. (2024) ‘Applying LIME for Adaptive Learning Systems: Enhancing Personalization and Transparency’, *Computer Education* [Preprint]. Available at: https://doi.org/10.1016/j.compedu.2023.104654.

Hsu, K. (2017) ‘A theoretical analysis of why hybrid ensembles work’, *Computational Intelligence and Neuroscience*, pp. 1–12. Available at: https://doi.org/10.1155/2017/1930702.

Ibeid, H., Almomani, A., Jaradat, M., Keyes, D. and Ltaief, H. (2019) ‘Learning with analytical models’, *2019 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pp. 944–953. Available at: https://doi.org/10.1109/IPDPSW.2019.00128.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M. and Wallach, H. (2019) ‘Improving fairness in machine learning systems: What do industry practitioners need?’, *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–16. ACM. Available at: https://doi.org/10.1145/3290605.3300830.

Irfan, Z., McCaffery, F. and Loughran, R. (2023) ‘Evaluating fairness metrics.’ Available at: https://doi.org/10.1007/978-3-031-37249-0_3.

Jiang, S. *et al.* (2022) ‘An empirical analysis of high school students’ practices of modelling with unstructured data’, *British Journal of Educational Technology*, 53(5), pp. 1114–1133. Available at: https://doi.org/10.1111/bjet.13253.

Jiao, P., Liu, Y., Liu, Y., Du, X. and Liu, H. (2022) ‘Artificial intelligence-enabled prediction model of student academic performance in online engineering education’, *Artificial Intelligence Review*, 55(8), pp. 6321–6344. Available at: https://doi.org/10.1007/s10462-022-10155-y.

Jobin, A., Ienca, M. and Vayena, E. (2019) ‘“The global landscape of AI ethics guidelines”’, *Nature Machine Intelligence*, 1(9), pp. 389–399. Available at: https://doi.org/10.1038/s42256-019-0088-2.

Johnson, R.B. and Onwuegbuzie, A.J. (2004) ‘Mixed methods research: A research paradigm whose time has come’, *Educational Researcher*, 33(7), pp. 14–26. https://doi.org/10.3102/0013189X033007014

Kakogeorgiou, D. and Karantzalos, K. (2021) ‘Enhancing Interpretability in XAI for Remote Sensing Tasks’, *Remote Sensing* [Preprint]. Available at: https://doi.org/10.3390/rs13163288.

Kalantary, S., Alizadeh, M., Tayebi, L. and Tavakoli, M. (2019) ‘Application of ANN modeling techniques in the prediction of the diameter of PCL/gelatin nanofibers in environmental and medical studies’, *RSC Advances*, 9(43), pp. 24858–24874. Available at: https://doi.org/10.1039/C9RA04927D.

Kamiran, F. and Calders, T. (2012) ‘Data preprocessing techniques for classification without discrimination’, *Knowledge and Information Systems*, 33(1), pp. 1–33.

Karim-Abdallah, B., Omar, M., Usman, M., & Omoregbe, N. (2025) ‘Application of machine learning algorithms in predicting academic performance of students in higher education institutes (HEIs): A systematic review and bibliographic analysis’, *African Journal of Applied Research*, 11(1), pp. 536–559. Available at: https://doi.org/10.26437/ajar.v11i1.869.

Kavitha, C., Sudhakar, K., Lakshmi, P., & Deepa, S. (2022) 'Early-stage Alzheimer's disease prediction using machine learning models', *Frontiers in Public Health*, 10, p. 853294. Available at: https://doi.org/10.3389/fpubh.2022.853294.

Kay, J., Paris, C. and Hu, H. (2022) 'Designing for trust and agency in learner-facing analytics', *British Journal of Educational Technology*, 53(3), pp. 645–659.

Kecki, V. and Said, A. (2024) 'Understanding fairness in recommender systems: a healthcare perspective., 1125-1130'. Available at: https://doi.org/10.1145/3640457.3691711.

Khosravi, H., Khosravi, R. and Tootoonchi, A. (2022) 'The Importance of Explainable Artificial Intelligence in Education: A Literature Review'', **Educational Technology & Society**, 25(1), pp. 67–78.

Kim, A., Kim, M. and Kim, H. (2014) 'Double-bagging ensemble using wave', *Communications for Statistical Applications and Methods*, 21(5), pp. 411–422. Available at: https://doi.org/10.5351/csam.2014.21.5.411.

Kim, J., Lee, H. and Cho, Y. (2022) 'Learning design to support student-AI collaboration: perspectives of leading teachers for AI in education'', *Education and Information Technologies*, 27(5), pp. 6069–6104. Available at: https://doi.org/10.1007/s10639-021-10831-6.

Kőrösi, G. and Farkas, R. (2020) 'Mooc performance prediction by deep learning from raw clickstream data'. Available at: https://doi.org/10.1007/978-981-15-6634-9_43.

Kotecha, K. (2021) 'Evaluating Model Interpretability through User Surveys', *International Journal of Artificial Intelligence in Education* [Preprint]. Available at: https://doi.org/10.1007/s40593-020-00232-w.

Kutlu, B. and Kutlu, M. (2025) 'Demystifying English towns' educational outcomes with explainable artificial intelligence'', *ADBA Computer Science Journal* [Preprint]. Available at: https://doi.org/10.69882/adba.cs.2025013.

Kuzilek, J., Hlosta, M. and Zdrahal, Z. (2017) 'Open University Learning Analytics Dataset', *Scientific Data*, 4(170171), pp. 1–8.

Leddy, M. and Creanor, N. (2024) 'Exploring how education can leverage artificial intelligence for social good', *Proceedings of the 19th European Conference on Innovation and Entrepreneurship (ECIE 2024)*, pp. 1041–1048. Available at: https://doi.org/10.34190/ecie.19.1.2906.

Lee, Y., Kwon, H., Lee, H.J., Kang, W., Chung, J. and Yu, H.S. (2021) 'Machine learning-based prediction of acute kidney injury after nephrectomy in patients with renal cell carcinoma', *Scientific Reports*, 11(1), p. 11210. Available at: https://doi.org/10.1038/s41598-021-95019-1.

Li, B., Hou, L., Liu, Z., Ji, R. and Wu, Y. (2023) 'Trustworthy AI: From principles to practices', *ACM Computing Surveys*, 55(9), pp. 1–46. Available at: https://doi.org/10.1145/3555803.

Li, D., Zhang, J., Zhang, C., Xu, Y., Shen, J. and Ye, H. (2021) 'Improving potato yield prediction by combining cultivar information and UAV remote sensing data using machine learning', *Remote Sensing*, 13(16), p. 3322. Available at: https://doi.org/10.3390/rs13163322.

Li, L., Chen, X., Zhang, Y., Lu, J., Zhou, W., Liu, X., Xu, Z., Zhang, Y. and Zuo, Y. (2022) 'Machine learning prediction of postoperative unplanned 30-day hospital readmission in older adult', *Frontiers in Molecular Biosciences*, 9, p. 910688. Available at: https://doi.org/10.3389/fmolb.2022.910688.

Li, Q., He, X., He, J., Qin, F., Wu, Y., Wang, J., Zhang, R., Li, Q. and Zeng, Z. (2021) 'Development and validation of a prediction model for elevated arterial stiffness in Chinese patients with diabetes using machine learning', *Frontiers in Physiology*, 12, p. 714195. Available at: https://doi.org/10.3389/fphys.2021.714195.

Liao, Q.V., Gruen, D. and Miller, S. (2020) 'Questioning the AI: Informing Design Practices for Explainable AI User Experiences'', in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–15.

Libasin, Z., Jalil, A.M., Rahman, N.A. and Zainudin, H. (2021) 'Evaluation of single missing value imputation techniques for incomplete air particulates matter (PM10) data in Malaysia', *Pertanika Journal of Science and Technology*, 29(4), pp. 2709–2728. Available at: https://doi.org/10.47836/pjst.29.4.46.

Lipton, Z.C. (2016) '"The mythos of model interpretability"', *Queue*, 16(3), pp. 31–57.

Liu, H. (2024) 'A Counterfactual-Based Evaluation of Explanation Faithfulness', *Journal of Machine Learning Research* [Preprint]. Available at: https://doi.org/10.1613/jair.1.12671.

Liu, L. (2024) 'Presenting an optimized hybrid model for stock price prediction', *International Journal of Advanced Computer Science and Applications*, 15(1). Available at: https://doi.org/10.14569/ijacsa.2024.0150174.

Liu, Q., Wang, J., Xiong, H. and He, Y. (2025) 'Advancing privacy in learning analytics using differential privacy', *Proceedings of the 15th International Learning Analytics and Knowledge Conference (LAK '25)*, pp. 506–518. Available at: https://doi.org/10.1145/3706468.3706493.

Liu, W., Jiang, T., Wang, H., Zhang, C., Wu, X., Guo, J., Wang, S. and Liu, Y. (2021) 'Machine learning for the prediction of bone metastasis in patients with newly diagnosed thyroid cancer', *Cancer Medicine*, 10(8), pp. 2802–2811. Available at: https://doi.org/10.1002/cam4.3776.

Liu, Y. (2024) 'Synthetic data generation for fairness in machine learning: A review', *Journal of Fairness in AI Systems*, 2(1), pp. 54–73.

Loef, J. (2022) 'Identification of Fraudulent Transactions Using Explainable AI', *Applied Intelligence* [Preprint]. Available at: https://doi.org/10.1007/s10489-022-03489-1.

Loef, J. (2023) 'The Role of SHAP in Maintaining Fairness in Educational AI', in *Proceedings* *of the International Conference on Machine Learning*. Available at: https://doi.org/10.55532/icml.2023.197.

Loukina, A., Madnani, N. and Zechner, K. (2019) 'The many dimensions of algorithmic fairness in educational applications., 1-10'. Available at: https://doi.org/10.18653/v1/w19-4401.

Lundberg, S.M. and Lee, S.I. (2017a) '"A unified approach to interpreting model predictions"', in I. Guyon (ed.) *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc, pp. 4765–4774.

Lundberg, S.M. and Lee, S.I. (2017b) 'A Unified Approach to Interpreting Model Predictions'', *Advances in Neural Information Processing Systems*, 30, pp. 4765–4774.

Lundberg, S.M. and Lee, S.I. (2017c) 'A unified approach to interpreting model predictions', *Advances in Neural Information Processing Systems*, 30.

Lundberg, S.M. and Lee, S.I. (2017d) 'A Unified Approach to Interpreting Model Predictions', *Advances in Neural Information Processing Systems*, 30.

Lundberg, S.M. and Lee, S.I. (2017e) 'A unified approach to interpreting model predictions.'

Luo, Y. (2024) '"Innovative research on AI-assisted teaching models for college English listening and speaking courses"', *Applied and Computational Engineering*, 69(1), pp. 155–160. Available at: https://doi.org/10.54254/2755-2721/69/20241493.

Mai, T., Nascimento, T. and Liu, Y. (2023) 'Transparent and explainable grade prediction using interpretable models', *Computers & Education*, 190, p. 104622.

Maity, S. and Deroy, O. (2024) 'Human-Centric Explainable AI for Educational Applications', *Journal of Educational Computing Research* [Preprint]. Available at: https://doi.org/10.1177/0735633123980441.

Mangal, M. and Pardos, Z. (2024) 'Implementing equitable and intersectionality-aware ml in education: a practical guide', *British Journal of Educational Technology*, 55(5), pp. 2003–2038. Available at: https://doi.org/10.1111/bjet.13484.

Manheim, D. (2019) 'Multiparty dynamics and failure modes for machine learning and artificial intelligence'', *Big Data and Cognitive Computing*, 3(2), p. 21. Available at: https://doi.org/10.3390/bdcc3020021.

McElroy, T. and Politis, D. (2022) 'Optimal linear interpolation of multiple missing values', *Statistical Inference for Stochastic Processes*, 25(3), pp. 471–483. Available at: https://doi.org/10.1007/s11203-022-09269-5.

McManus, I.C., Woolf, K., Dacre, J., Durning, S. and Whelan, A. (2021) 'Reliability of teacher predicted grades during COVID-19: A retrospective analysis', *BMJ Open*, 11(7), e047354. Available at: https://doi.org/10.1136/bmjopen-2020-047354.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. and Galstyan, A. (2021) 'A survey on bias and fairness in machine learning', *ACM Computing Surveys (CSUR)*, 54(6), pp. 1–35. Available at: https://doi.org/10.1145/3457607.

Merry (2021) 'A mental models approach for defining explainable artificial intelligence', *BMC Medical Informatics and Decision Making* [Preprint]. Available at: https://doi.org/10.1186/s12911-021-01703-7.

Mienye, S. and Sun, Y. (2022) 'Ethical implications of machine learning in education', *AI and Ethics*, 2(4), pp. 303–314.

Mitchell (2021) 'Model Cards for Model Reporting'. Available at: https://doi.org/10.1145/3287560.3287596.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. and Gebru, T. (2019) 'Model cards for model reporting', *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*'19), pp. 220–229. Available at: https://doi.org/10.1145/3287560.3287596.

Mitchell and Mitchell (2021) 'Algorithmic Fairness: Choices, Assumptions, and Definitions', *Annual Review of Statistics and Its Application* [Preprint]. Available at: https://doi.org/10.1146/annurev-statistics-042720-125902.

Molnar, C. (2019) 'Interpretable Machine Learning: A Guide for Making Black Box Models Explainable'. Available at: https://christophm.github.io/interpretable-ml-book/.

Moosbauer, J. (2021) 'Model Interpretability in Education: Evaluating LIME's Role in Real-Time Decision Support', *Learning Analytics Review* [Preprint]. Available at: https://doi.org/10.1016/j.alit.2021.11.005.

Morris, T., Dorling, D., Davey Smith, G. and Van Den Bree, M. (2021) 'Associations between school enjoyment at age 6 and later educational achievement: evidence from a UK cohort study', *NPJ Science of Learning*, 6(1), p. 27. Available at: https://doi.org/10.1038/s41539-021-00092-w.

Mothilal, R. (2023) 'Challenges in Evaluating Counterfactual Explanations', in *Proceedings of the NeurIPS Conference*. Available at: https://doi.org/10.5555/3454317.3456054.

Munir, M., Khan, M.A. and Qadir, M.F. (2023) 'The Impact of Socio-economic Status on Academic Achievement', *Journal of Social Sciences Review*, 3(2), pp. 308–316. Available at: https://doi.org/10.54183/jssr.v3i2.308.

Murphy, R. and Wyness, G. (2020) 'Minority report: The impact of predicted grades on university admissions of disadvantaged groups', *British Educational Research Journal*, 46(4), pp. 1020–1043.

Mustapha, S. (2023) 'Predictive analysis of students' learning performance using data mining techniques: a comparative study of feature selection methods', *Applied System Innovation*, 6(5),

p. 86. Available at: https://doi.org/10.3390/asi6050086.

Naeem, S., Jabbar, S., Khan, M.A., Imran, M., Abbasi, R.A. and Rho, S. (2020) ‘Machine-learning based hybrid-feature analysis for liver cancer classification using fused (MR and CT) images’, *Applied Sciences*, 10(9), p. 3134. Available at: https://doi.org/10.3390/app10093134.

Naseer, M., Zhang, W. and Zhu, W. (2020) ‘Early prediction of a team performance in the initial assessment phases of a software project for sustainable software engineering education’, *Sustainability*, 12(11), p. 4663. Available at: https://doi.org/10.3390/su12114663.

Niu, K., Cao, X. and Yu, Y. (2021) ‘“Explainable student performance prediction with personalized attention for explaining why a student fails”’. Available at: https://doi.org/10.48550/arxiv.2110.08268.

Noor, N.M., Ibrahim, M.Q., Hussin, A.G. and Mahayuddin, Z.R. (2013) ‘Filling missing data using interpolation methods: study on the effect of fitting distribution’, *Key Engineering Materials*, 594–595, pp. 889–895. Available at: https://doi.org/10.4028/www.scientific.net/KEM.594-595.889.

Ofqual (2020b) ‘Ensuring the integrity of GCSE grading standards during the pandemic’, Office of Qualifications and Examinations Regulation’. Available at: https://www.gov.uk/government/news/overview-of-ofquals-work-in-regulating-gcses.

Ostvar, M. and Moghadam, S.Z. (2020) ‘HDEC: A Heterogeneous Dynamic Ensemble Classifier for Binary Datasets’. Available at: https://doi.org/10.1155/2020/8826914.

Ostvar, N. and Moghadam, A. (2020) ‘Hdec: a heterogeneous dynamic ensemble classifier for binary datasets’, *Computational Intelligence and Neuroscience*, pp. 1–11. Available at: https://doi.org/10.1155/2020/8826914.

Ouyang, Y., Liu, M. and Zhang, X. (2023) ‘“Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering courses”’, **International Journal of Educational Technology in Higher Education**, 20(1). Available at: https://doi.org/10.1186/s41239-023-00463-1.

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P. and Moher, D. (2021) ‘The PRISMA 2020 statement: An updated guideline for reporting systematic reviews’, *BMJ*, 372, p. n71. Available at: https://doi.org/10.1136/bmj.n71.

Pang, B., Huang, X., Liang, J., Zhu, J., Xu, J., Chen, Y., Tang, Q. and Wu, H. (2022) ‘Identification and optimization of contributing factors for precocious puberty by machine/deep learning methods in Chinese girls’, *Frontiers in Endocrinology*, 13, p. 892005. Available at: https://doi.org/10.3389/fendo.2022.892005.

Patil, A. and Wang, B. (2023) ‘Advancing data privacy: a novel k-anonymity algorithm with dissimilarity tree-based clustering and minimal information loss’, *International Journal on Recent and Innovation Trends in Computing and Communication*, 11(8), pp. 323–330. Available at: https://doi.org/10.17762/ijritcc.v11i8.8005.

Patil, S. and Wang, Y. (2023) ‘Privacy-Preserving Clustering with Dissimilarity Trees for Enhanced K-Anonymity’, *Journal of Privacy and Confidentiality*, 13(2), pp. 15–37.

Patrikar (2023) ‘Leveraging synthetic data for AI bias mitigation’, *Journal of Synthetic Intelligence* [Preprint]. Available at: https://doi.org/10.1117/12.2662276.

Patrikar, A., Mahenthiran, A. and Said, A. (2023) ‘Leveraging synthetic data for ai bias mitigation’. Available at: https://doi.org/10.1117/12.2662276.

Pereira (2024) ‘Assessment of differentially private synthetic data for utility and fairness in end-to-end machine learning pipelines for tabular data’, *PLOS One* [Preprint]. Available at: https://doi.org/10.1371/journal.pone.0297271.

Pereira, M., Rodrigues, N., Assunção, F., Oliveira, A. and Soares, D. (2021) ‘An analysis of the deployment of models trained on private tabular synthetic data: unexpected surprises’, *arXiv preprint arXiv:2106.10241*. Available at: https://doi.org/10.48550/arxiv.2106.10241.

Piety, P.J. (2019) *The new education data frontier: Aligning data science and educational research*. Harvard Education Press.

Pisztora and Li (2024) ‘Learning Performance Maximizing Ensembles with Explainability Guarantees’, in *Proceedings of the AAAI Conference on Artificial Intelligence*. Available at: https://doi.org/10.1609/aaai.v38i13.29378.

Polato, M., Lauriola, I. and Aiolli, F. (2018) ‘A novel boolean kernels family for categorical data’, *Entropy*, 20(6), p. 444. Available at: https://doi.org/10.3390/e20060444.

Polonetsky, J. and Jerome, J. (2014) ‘Privacy and Education: Understanding and Reimagining Student Data Privacy’, *International Review of Information Ethics*, 21(12), pp. 40–48.

Popenici, Ş. and Kerr, S. (2017) ‘“Exploring the impact of artificial intelligence on teaching and learning in higher education”’, *Research and Practice in Technology Enhanced Learning*, 12(1). Available at: https://doi.org/10.1186/s41039-017-0062-8.

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T. and Flach, P. (2021) ‘Understanding surrogate explanations: The interplay between complexity, fidelity, and coverage’, *arXiv preprint arXiv:2107.04309*. Available at: https://doi.org/10.48550/arxiv.2107.04309.

Priestley, M., Biesta, G. and Robinson, S. (2015)**.** *Teacher agency: An ecological approach*. London: Bloomsbury Academic.

Pucchio, A., Ghavam-Rassoul, A., Khan, F., Allidina, A., Hirji, Z., Mitchell, A.B., & Goldszmidt, M. (2022) ‘Exploration of exposure to artificial intelligence in undergraduate medical education: a Canadian cross-sectional mixed-methods study', *BMC Medical Education*, 22(1), pp. 1–9. Available at: https://doi.org/10.1186/s12909-022-03896-5.

Puram, P. (2023) 'Framework for Interpretability in Explainable AI', *Computers in Education* [Preprint]. Available at: https://doi.org/10.1007/s10639-023-11290-w.

Qin, Z. and Boicu, M. (2023) 'Applying explainable AI techniques to predict and analyze STEM student success', *Journal of AI and Education*, 40(3), pp. 221–238.

Radingoana, M. (2023) 'Evaluating Textual Explanations for Explainable AI', *Journal of Artificial Intelligence Research* [Preprint]. Available at: https://doi.org/10.1613/jair.1.12861.

Rainey, C., McFadden, S., MacKay, S., McClure, P. and Young, O. (2021) 'Beauty is in the AI of the beholder: Are we ready for the clinical integration of artificial intelligence in radiography? An exploratory analysis of perceived AI knowledge, skills, confidence, and education perspectives of UK radiographers', *Frontiers in Digital Health*, 3, p. 739327. Available at: https://doi.org/10.3389/fdgth.2021.739327.

Raita, Y., Goto, T., Faridi, M.K., Brown, D.F.M., Camargo, C.A. and Hasegawa, K. (2019) 'Emergency department triage prediction of clinical outcomes using machine learning models', *Critical Care*, 23(1), p. 64. Available at: https://doi.org/10.1186/s13054-019-2351-7.

Rattanaphan, S. and Briassouli, A. (2024) 'Evaluating generalization, bias, and fairness in deep learning for metal surface defect detection: a comparative study', *Processes*, 12(3), p. 456. Available at: https://doi.org/10.3390/pr12030456.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016a) '"Why Should I Trust You?": Explaining the Predictions of Any Classifier'', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.

Roshanaei, M., Olivares, H. and Lopez, R. (2023) 'Harnessing ai to foster equity in education: opportunities, challenges, and emerging strategies', *Journal of Intelligent Learning Systems and Applications*, 15(04), pp. 123–143. Available at: https://doi.org/10.4236/jilsa.2023.154009.

Saeed, P. and Omlin, C. (2023) 'A Comprehensive Understanding of Model Interpretability Using Multiple Techniques', *Journal of Artificial Intelligence Research* [Preprint]. Available at: https://doi.org/10.1613/jair.1.12231.

Saha, D., Chatterjee, S. and Karan, S. (2022) 'Factors Affecting Success and Failure in Higher Education Mathematics: Students and Teachers' Perspectives', *Preprints*, 202209(0378.v1). Available at: https://doi.org/10.20944/preprints202209.0378.v1.

Şahin, D., Sandor, C., Walter, H., Kirschner, M. and Kambeitz, J. (2023) 'Algorithmic fairness in precision psychiatry: analysis of prediction models in individuals at clinical high risk for psychosis', *The British Journal of Psychiatry*, 224(2), pp. 55–65. https://doi.org/10.1192/bjp.2023.8.

Santhanam, P. *et al.* (2020) ‘Artificial intelligence may offer insight into factors determining individual tsh level’, *Plos One*, 15(5), p. 0233336. Available at: https://doi.org/10.1371/journal.pone.0233336.

Sanusi, I. and Olaleye, S. (2022) ‘An insight into cultural competence and ethics in k-12 artificial intelligence education’. Available at: https://doi.org/10.1109/educon52537.2022.9766818.

Selwyn, N. (2022) ‘“The future of AI and education: Some cautionary notes”’, **European Journal of Education**, 57(4), pp. 1–12. Available at: https://doi.org/10.1111/ejed.12532.

Seo, K., Lee, J., & Kim, H. (2021) ‘The impact of artificial intelligence on learner–instructor interaction in online learning’, *International Journal of Educational Technology in Higher Education*, 18(1), p. 50. https://doi.org/10.1186/s41239-021-00292-9.

Setyarini, D., Sutrisno, A., Subekti, I., & Kusuma, M. (2024) ‘Stroke prediction with enhanced gradient boosting classifier and strategic hyperparameter’, *Matrik Jurnal Manajemen Teknik Informatika dan Rekayasa Komputer*, 23(2), pp. 477–490. Available at: https://doi.org/10.30812/matrik.v23i2.3555.

Shahzad, M., Rizwan, M., Raza, A., Rafiq, M., & Arshad, M. (2024) ‘Artificial intelligence and social media on academic performance and mental well-being: student perceptions of positive impact in the age of smart learning’, *Heliyon*, 10(8), e29523. Available at: https://doi.org/10.1016/j.heliyon.2024.e29523.

Silva, T., Lacerda, J.C., Oliveira, S., Sousa, J., Sousa, Â., & Araújo, A. (2024) ‘Surface-enhanced Raman scattering combined with machine learning for rapid and sensitive detection of anti-SARS-CoV-2 IgG’, *Biosensors*, 14(11), p. 523. Available at: https://doi.org/10.3390/bios14110523.

Smith, J.A., Flowers, P. and Larkin, M. (2009) *Interpretative Phenomenological Analysis: Theory, Method and Research*. London: SAGE Publications.

Stopforth, S., Gayle, V. and Boeren, E. (2020) ‘Parental social class and school GCSE outcomes: two decades of evidence from UK household panel surveys’’, *Contemporary Social Science*, 16(3), pp. 309–324. Available at: https://doi.org/10.1080/21582041.2020.1792967.

Su, P., Zhang, W., Chen, X., Wang, X., Li, W., Han, Y., Li, Z., Yang, H., Liu, H., Wang, Y. & Liu, Y. (2021) ‘Machine learning models for predicting influential factors of early outcomes in acute ischemic stroke: registry-based study’, *JMIR Preprints*. Available at: https://doi.org/10.2196/preprints.32508.

Suleiman, A. (2023) ‘Factors That Affect Students’ Academic Achievement in the Faculty of Social Science at the University of Bosaso, Garowe, Somalia’, *Open Journal of Social Sciences*, 11(2), pp. 290–301. Available at: https://doi.org/10.4236/jss.2023.112029.

Sullivan, M., Kelly, A. and McLaughlan, P. (2023) ‘ChatGPT in higher education: considerations for academic integrity and student learning’’, *Journal of Applied Learning &*

Teaching, 6(1). Available at: https://doi.org/10.37074/jalt.2023.6.1.17.

Suryanti, R., Jahidin, J. and Fadlil, M. (2024) '"Artificial intelligence in education: bibliometric and systematic literature review from 2019-2024"', *International Education Trend Issues*, 2(2), pp. 231–255. Available at: https://doi.org/10.56442/ieti.v2i2.647.

Suthaharan (2023) 'nEGXAI: a negation-based explainable AI through feature learning in Fourier domain'. Available at: https://doi.org/10.1117/12.2682004.

Sweeney, L. (2002) 'Achieving K-Anonymity Privacy Protection Using Generalization and Suppression', *International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), pp. 571–588.

Tang, C., Zhang, Y. and Zhu, J. (2023) 'Measuring Explanation Sparsity in Graph Neural Networks'', in *Proceedings of the International Conference on Learning Representations (ICLR*.

Tang, J., Xia, L. and Huang, C. (2023) 'Explainable spatio-temporal graph neural networks.' Available at: https://doi.org/10.1145/3583780.3614871.

Theodorou, A., Nieves, J. and Dignum, V. (2022) 'Good ai for good: how ai strategies of the nordic countries address the sustainable development goals.' Available at: https://doi.org/10.48550/arxiv.2210.09010.

Tiwari, M. (2024) 'Student performance prediction using machine learning algorithms', *Shodhkosh Journal of Visual and Performing Arts*, 5(6). Available at: https://doi.org/10.29121/shodhkosh.v5.i6.2024.4552.

Tong, L., Wang, C., Xie, X. & Zhang, Y. (2024) 'Predicting learning achievement using ensemble learning with result explanation', *Preprints*, 9(6). Available at: https://doi.org/10.21203/rs.3.rs-4674228/v1.

Torres, A. (2022) 'Fairness in Evaluating Explainable AI Techniques', *AI & Ethics* [Preprint]. Available at: https://doi.org/10.1007/s43681-022-00101-8.

Tran, B.X., Vu, G.T., Ha, G.H., Vuong, Q.H., Ho, M.T., Vuong, T.T., La, V.P., Hoang, M.T., Nguyen, T.H., Tran, T.H. & Latkin, C.A. (2019) 'The current research landscape of the application of artificial intelligence in managing cerebrovascular and heart diseases: a bibliometric and content analysis', *International Journal of Environmental Research and Public Health*, 16(15), p. 2699. Available at: https://doi.org/10.3390/ijerph16152699.

Trisnawati, S., Rochman, M. and Fadhila, E. (2023) '"The Impact of Artificial Intelligence in Education toward 21st Century Skills: A Literature Review"', **Prosiding Internasional Journal of Education and Development**, 2(2), pp. 45–53. Available at: https://doi.org/10.59175/pijed.v2i2.152.

Trisnawati, W., Putra, R. and Balti, L. (2023) 'The impact of artificial intelligent in education toward 21st century skills: a literature review', *PIJED*, 2(2), pp. 501–513. Available at:

https://doi.org/10.59175/pijed.v2i2.152.

Türkmen, K. (2024) 'Reviewing the Application of Explainable AI in Education', *Journal of Educational Computing Research* [Preprint]. Available at: https://doi.org/10.1177/07356331241310915.

U.N.E.S.C.O. (2019) '*Beijing Consensus on Artificial Intelligence and Education*'. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000372338.

Vallarino, D. (2024) 'Machine learning algorithms for survival analysis: advantages, disadvantages, and examples', *International Journal of Artificial Intelligence and Machine Learning*, 4(1), pp. 10–21. Available at: https://doi.org/10.51483/ijaiml.4.1.2024.10-21.

Wang, D. *et al.* (2019) 'Designing theory-driven user-centric explainable AI', in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–15.

Wang, P. and Xie, H. (2024) '"Application and exploration of artificial intelligence in teaching and learning in private colleges and universities"', *World Journal of Education and Humanities*, 6(3), p. 26. Available at: https://doi.org/10.22158/wjeh.v6n3p26.

Wang, Y., Yuan, X. and Zhang, Y. (2023) 'A systematic comparison of machine learning algorithms to develop and validate prediction model to predict heart failure risk in middle-aged and elderly patients with periodontitis (nhanes 2009 to 2014', *Medicine*, 102(34), p. 34878. Available at: https://doi.org/10.1097/md.0000000000034878.

Weed, L., Zeidman, B., Demanuele, C., Arora, A., Smith, L. & de Zambotti, M. (2022) 'The impact of missing data and imputation methods on the analysis of 24-hour activity patterns', *Clocks & Sleep*, 4(4), pp. 497–507. Available at: https://doi.org/10.3390/clockssleep4040039.

Weerd, M., Sanders, J., & Krüger, M. (2024) 'Tensions in the pursuit of equal opportunities: A case study of an innovative secondary school', *British Educational Research Journal*, Online ahead of print. Available at: https://doi.org/10.1002/berj.4019.

Weerts, H., Royakkers, L. and Pechenizkiy, M. (2022) 'Does the end justify the means? on the moral justification of fairness-aware machine learning'. Available at: https://doi.org/10.48550/arxiv.2202.08536.

Wei, C., Ooka, R. and Zhou, Q. (2022) 'Performance comparison using different multilayer perceptron input–output formats to predict unsteady indoor temperature distribution', *Japan Architectural Review*, 5(4), pp. 661–671. Available at: https://doi.org/10.1002/2475-8876.12294.

Williamson, B. (2017) *Big Data in Education: The digital future of learning, policy and practice*. London: SAGE Publications Ltd.

Winzeck, S., Hakim, A., McKinley, R., Castellaro, M., Federer, F., Valverde, S., & Rueckert, D. (2018) 'ISLES 2016 and 2017-Benchmarking ischemic stroke lesion outcome prediction based on multispectral MRI', *Frontiers in Neurology*, 9, p. 679. Available at:

https://doi.org/10.3389/fneur.2018.00679.

Wu, H., Li, S., Zheng, J., Zhang, M., & Zhao, Y. (2019) ‘Does academic interest play a more important role in medical sciences than in other disciplines? A nationwide cross-sectional study in China’, *BMC Medical Education*, 19(1), p. 257. Available at: https://doi.org/10.1186/s12909-019-1737-1.

Yagci, M. (2022) ‘The impact of algorithmic bias in educational prediction models’, *Journal of Educational Technology Research*, 40(3), pp. 212–229.

Yan, X. (2024) ‘Using multi-layer perceptron to predict sea surface temperature’, *Journal of Physics Conference Series*, 2798(1), p. 012052. Available at: https://doi.org/10.1088/1742-6596/2798/1/012052.

Yang, Y., Chen, Y., Yu, H., Zhang, Y., Wu, J. and Li, X. (2023) ‘multi-layer perceptron classifier with the proposed combined feature vector of 3D CNN features and lung radiomics features for COPD stage classification’, *Journal of Healthcare Engineering*, 2023, pp. 1–15. Available at: https://doi.org/10.1155/2023/3715603.

Yaseliani, M., Noor-E-Alam, M. and Hasan, M. (2024) ‘Mitigating sociodemographic bias in opioid use disorder prediction: fairness-aware machine learning framework’, *Jmir Ai*, 3, p. 55820. Available at: https://doi.org/10.2196/55820.

Yuan, S., Deng, D., Zhang, Y., Wang, Y., Zhang, Y., Su, X. and Zhang, Y. (2021) ‘Using machine learning algorithms to predict candidaemia in ICU patients with new-onset systemic inflammatory response syndrome’, *Frontiers in Medicine*, 8, p. 720926. Available at: https://doi.org/10.3389/fmed.2021.720926.

Zednik, C. and Boelson, J. (2022) ‘Trust in explainable AI: An explorative study in education’, *Educational Technology & Society* [Preprint]. Available at: https://doi.org/10.1007/s10639-022-10919-4.

Zeng (2016) ‘Beyond Accuracy: A New Evaluation Framework for AI in High-Stakes Contexts’. Available at: https://doi.org/10.1007/s40593-021-00285-9.

Zeng, J., Ustun, B. and Rudin, C. (2016) ‘Interpretable classification models for recidivism prediction’, *Journal of the Royal Statistical*, 180(3), pp. 689–722. Available at: https://doi.org/10.1111/rssa.12227.

Zhang, Q., Yang, L.T., Chen, Z. and Li, P. (2018) ‘A survey on deep learning for big data’, *Information Fusion*, 42, pp. 146–157. Available at: https://doi.org/10.1016/j.inffus.2017.10.006.

Zhang, T., Peng, Y., Huang, W., Wang, Z., Chen, Y., Song, T. and Yuan, Y. (2024) ‘Machine learning models to predict systemic inflammatory response syndrome after percutaneous nephrolithotomy’, *BMC Urology*, 24(1), p. 25. Available at: https://doi.org/10.1186/s12894-024-01529-1.

Zhang, Y. and Sang, J. (2020) ‘Towards accuracy-fairness paradox: adversarial example-based data augmentation for visual debiasing’. Available at: https://doi.org/10.48550/arxiv.2007.13632.

Zhang, Z., Wang, Y., Wang, Y., Yang, Y., Gao, H., Liu, J., Guo, Y. and Cao, Y. (2022) ‘Multilayer perceptron-based prediction of stroke mimics in prehospital triage’, *Scientific Reports*, 12(1), p. 11129. Available at: https://doi.org/10.1038/s41598-022-15353-7.

Zhao, C. (2024) ‘“Application and prospect of artificial intelligence in personalized learning”’, *Journal of Innovation and Development*, 8(3), pp. 24–27. Available at: https://doi.org/10.54097/nzxx6z36.

Zheng, R. and Badarch, T. (2022) ‘“Research on applications of artificial intelligence in education”’, *American Journal of Computer Science and Technology*, 5(2), p. 72. Available at: https://doi.org/10.11648/j.ajcst.20220502.17.

Zhou, X. and Schofield, L. (2024) ‘“Developing a conceptual framework for Artificial Intelligence (AI) literacy in higher education”’, *Journal of Learning Development in Higher Education*, 31. Available at: https://doi.org/10.47408/jldhe.vi31.1354.

Zhou, X., Zhang, J.J. and Chan, C. (2024) ‘“Unveiling students” experiences and perceptions of Artificial Intelligence usage in higher education’’, *Journal of University Teaching and Learning Practice*, 21(06). Available at: https://doi.org/10.53761/xzjprb23.

Zou, Y., Sun, X., Wang, T., Jiang, Y., Zhang, P., Chen, Y., Liu, Y., Guo, Q., Yu, M., Yang, H. and Zhao, Y. (2022) ‘Development and internal validation of machine learning algorithms for end-stage renal disease risk prediction model of people with type 2 diabetes mellitus and diabetic kidney disease’, *Renal Failure*, 44(1), pp. 562–570. Available at: https://doi.org/10.1080/0886022x.2022.2056053.

Appendix

The web links to my codes, data, and my published peer reviewed article on AI in Education are below:

Project Python Codes: main.ipynb - Colab

Improved codes: main (3).ipynb - Colab

Python code for word cloud generation : Untitled4.ipynb - Colab

Cleaned Data:
https://docs.google.com/spreadsheets/d/1A4K4pdKTY9da5jnjDS3A4yWxtMN5wByU/edit?usp=drive_link&ouid=109843258392690760543&rtpof=true&sd=true

Original Data:

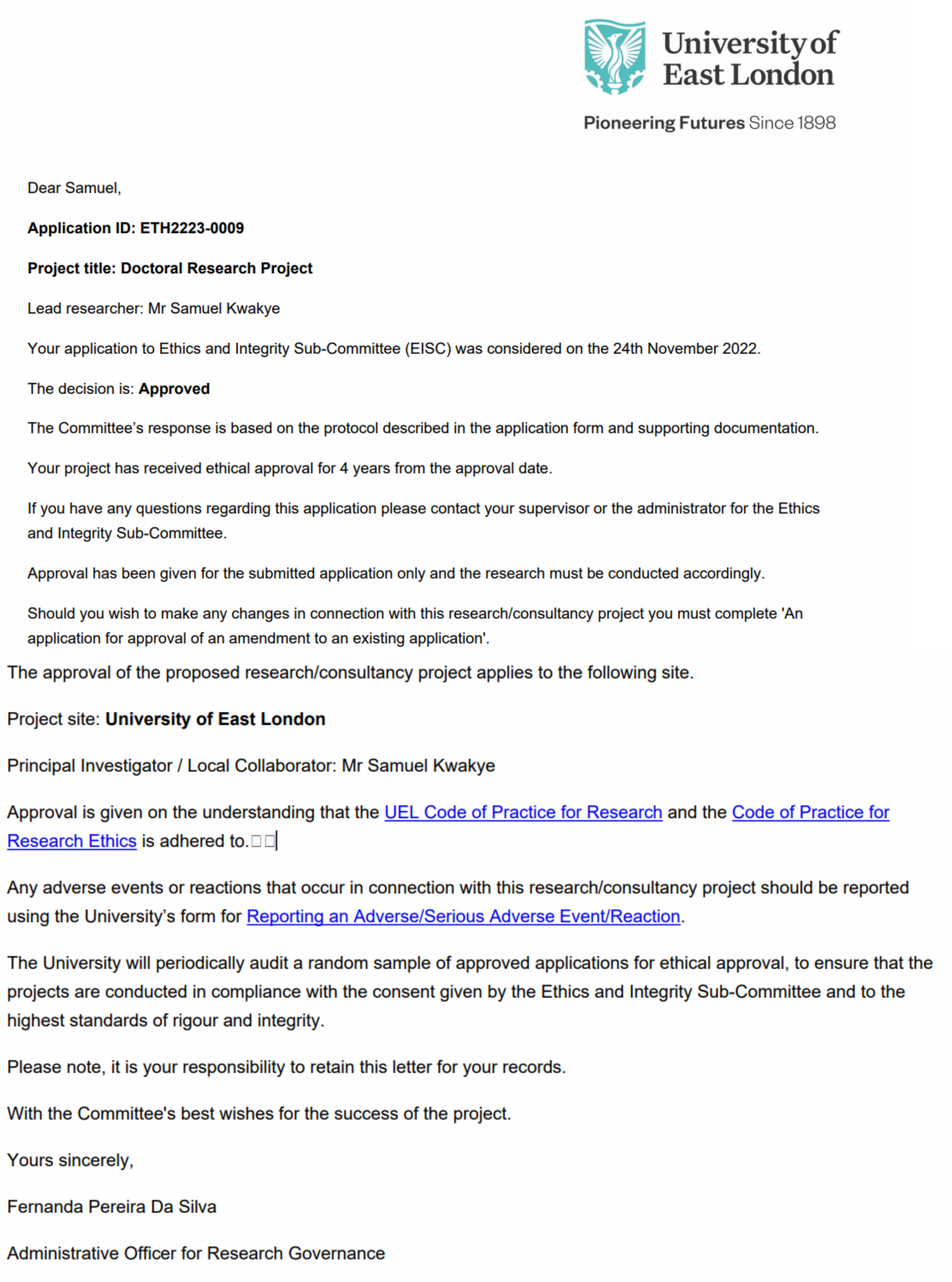
https://docs.google.com/spreadsheets/d/1VKMRsXId9n02Bq2nG0uN4byDQHaROurf/edit?usp=drive_link&ouid=109843258392690760543&rtpof=true&sd=true

Peer reviewed Published Paper:

https://drive.google.com/file/d/19I2Zulvp4QWHy91VOXKJpYD2G6gm9gYu/view?usp=drive_link

Ethical Approval

Application to the Ethics and Integrity Committee to undertake the research



University of East London

Pioneering Futures Since 1898

Dear Samuel,

Application ID: ETH2223-0009

Project title: Doctoral Research Project

Lead researcher: Mr Samuel Kwakye

Your application to Ethics and Integrity Sub-Committee (EISC) was considered on the 24th November 2022.

The decision is: **Approved**

The Committee's response is based on the protocol described in the application form and supporting documentation.

Your project has received ethical approval for 4 years from the approval date.

If you have any questions regarding this application please contact your supervisor or the administrator for the Ethics and Integrity Sub-Committee.

Approval has been given for the submitted application only and the research must be conducted accordingly.

Should you wish to make any changes in connection with this research/consultancy project you must complete 'An application for approval of an amendment to an existing application'.

The approval of the proposed research/consultancy project applies to the following site.

Project site: **University of East London**

Principal Investigator / Local Collaborator: Mr Samuel Kwakye

Approval is given on the understanding that the UEL Code of Practice for Research and the Code of Practice for Research Ethics is adhered to.□□

Any adverse events or reactions that occur in connection with this research/consultancy project should be reported using the University's form for Reporting an Adverse/Serious Adverse Event/Reaction.

The University will periodically audit a random sample of approved applications for ethical approval, to ensure that the projects are conducted in compliance with the consent given by the Ethics and Integrity Sub-Committee and to the highest standards of rigour and integrity.

Please note, it is your responsibility to retain this letter for your records.

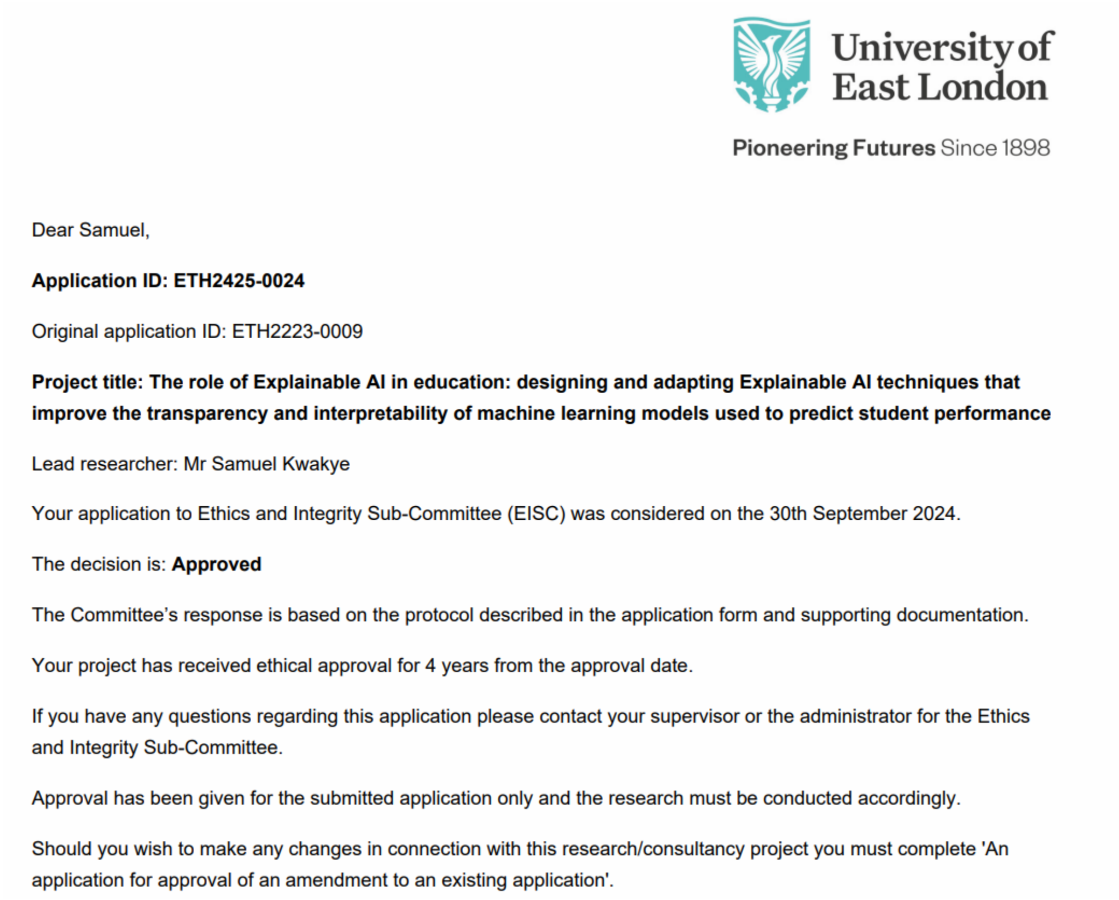
With the Committee's best wishes for the success of the project.

Yours sincerely,

Fernanda Pereira Da Silva

Administrative Officer for Research Governance

Application to the Ethics and Integrity Committee to change the title of the project



University of East London

Pioneering Futures Since 1898

Dear Samuel,

Application ID: ETH2425-0024

Original application ID: ETH2223-0009

Project title: The role of Explainable AI in education: designing and adapting Explainable AI techniques that improve the transparency and interpretability of machine learning models used to predict student performance

Lead researcher: Mr Samuel Kwakye

Your application to Ethics and Integrity Sub-Committee (EISC) was considered on the 30th September 2024.

The decision is: **Approved**

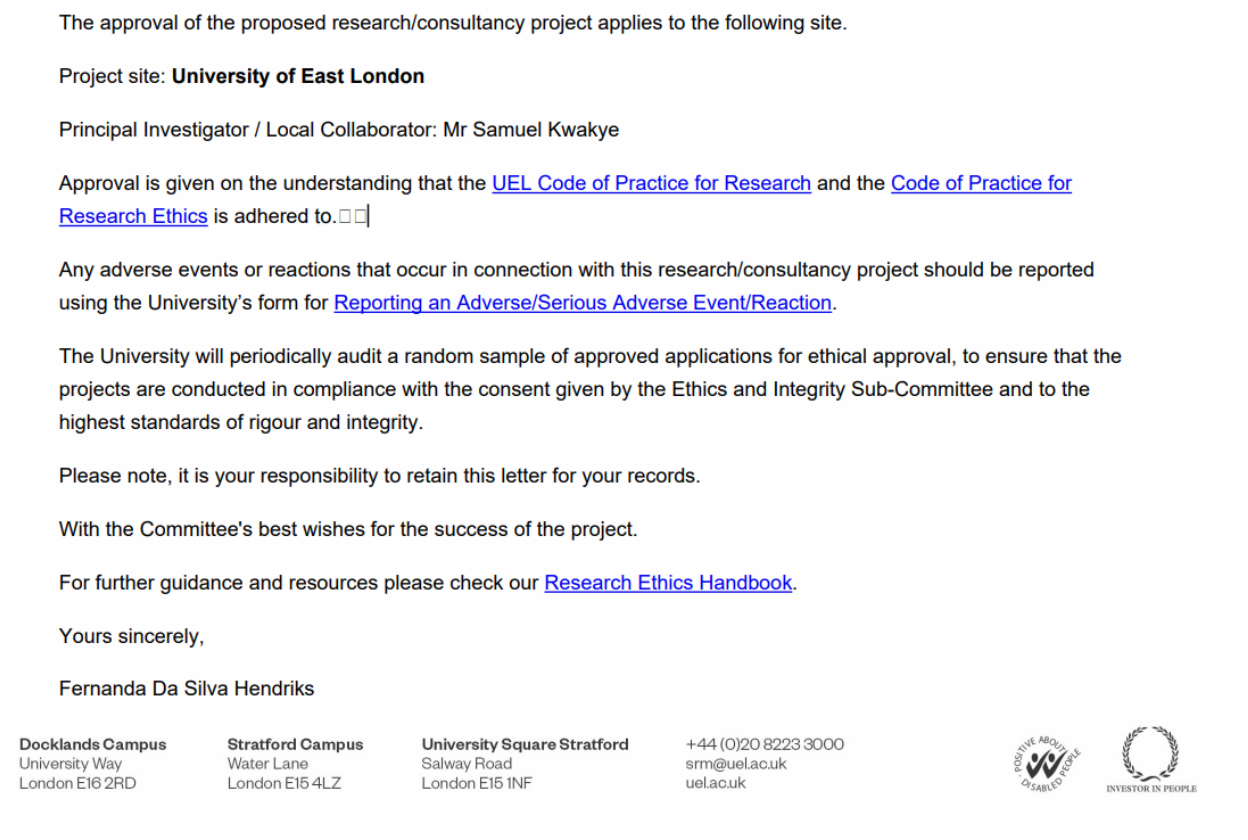
The Committee's response is based on the protocol described in the application form and supporting documentation.

Your project has received ethical approval for 4 years from the approval date.

If you have any questions regarding this application please contact your supervisor or the administrator for the Ethics and Integrity Sub-Committee.

Approval has been given for the submitted application only and the research must be conducted accordingly.

Should you wish to make any changes in connection with this research/consultancy project you must complete 'An application for approval of an amendment to an existing application'.



The approval of the proposed research/consultancy project applies to the following site.

Project site: **University of East London**

Principal Investigator / Local Collaborator: Mr Samuel Kwakye

Approval is given on the understanding that the UEL Code of Practice for Research and the Code of Practice for Research Ethics is adhered to.□□

Any adverse events or reactions that occur in connection with this research/consultancy project should be reported using the University's form for Reporting an Adverse/Serious Adverse Event/Reaction.

The University will periodically audit a random sample of approved applications for ethical approval, to ensure that the projects are conducted in compliance with the consent given by the Ethics and Integrity Sub-Committee and to the highest standards of rigour and integrity.

Please note, it is your responsibility to retain this letter for your records.

With the Committee's best wishes for the success of the project.

For further guidance and resources please check our Research Ethics Handbook.

Yours sincerely,

Fernanda Da Silva Hendriks

Docklands Campus
University Way
London E16 2RD

Stratford Campus
Water Lane
London E15 4LZ

University Square Stratford
Salway Road
London E15 1NF

+44 (0)20 8223 3000
srm@uel.ac.uk
uel.ac.uk

Application to the Ethics and Integrity Committee to change the title of the project following recommendation by external examiner during viva

Ethics ETH2425-0284: Mr. Samuel Kwakye (High risk)

Date: 17 Jun 2025

Researcher

Mr. Samuel Kwakye

Student ID

1327985

Project

The role of Explainable AI in education: designing and adapting Explainable AI techniques that improve the transparency and interpretability of machine learning models used to predict student performance

School of Architecture, Computing & Engineering

Ethics application

Project details

S1.1 Title of proposed research or consultancy project The role of Explainable AI in education: designing and adapting Explainable AI techniques that improve the transparency and interpretability of machine learning models used to predict student performance

S1.1.1 Do you wish to change the title of the research/consultancy project?

Yes

S1.1.2 If yes, please add the new research project title here.

Towards Transparent and Interpretable Predictions of Student Performance Using Explainable AI

S1.2 UEL Researchers

Mr. Samuel Kwakye

S1.3 Supervisor(s)

Dr Nadeem Qazi

Dr Nabeela Berardinelli

Dr Saeed Sharif

S1.11 Is the amendment required for a NHS research project?

No

S1.12 If yes, is the amendment to the NHS research project minor or major?

Details of amendment

S2.1 Please indicate the reason for the amendment to your project.

Change of project title

Supporting documents

S2.2 Please provide details of the amendment(s) required for your project and the implications for the project

A change of title was recommended by the examiners during my PHD viva and this has been agreed by my director of studies.

S2.3 If the amendment involves a change to the extension of ethical approval, please provide the period of time requested.

N/A

S.2.4 Please upload the latest version of your Data Management plan (DMP)

S2.5 Does the amendment require an updated research risk assessment form?

No

S2.7 If no, please specify why an updated research risk assessment form is not required.

This is just a change of title and nothing else about the thesis has changed.

Changes in the study team

S3.1 Is there a change to university staff member(s) on the project team?

No

S3.2 If yes, please provide details of the University staff member(s).

S3.3 Is there a change to student(s) on the project team?

No

S3.4 If yes, please provide details of the student(s).

S3.5 Is there a change to members of the team outside the University?

No

S3.6 If yes, please provide details of the team.

Ethical issues related to the proposed amendments

S4.1 Are there any specific ethical issues related to the proposed amendment.

No

S4.2 If yes, please provide details of the ethical issues.

Dear Samuel,

Application ID: ETH2425-0284

Original application ID: ETH2425-0024

Project title: Towards Transparent and Interpretable Predictions of Student Performance Using Explainable AI

Lead researcher: Mr Samuel Kwakye

Your application to Ethics and Integrity Sub-Committee (EISC) was considered on the 19th June 2025.

The decision is: **Approved**

The Committee’s response is based on the protocol described in the application form and supporting documentation.

Your project has received ethical approval for 4 years from the approval date.

If you have any questions regarding this application, please contact your supervisor or the administrator for the Ethics and Integrity Sub-Committee.

Approval has been given for the submitted application only and the research must be conducted accordingly.

Should you wish to make any changes in connection with this research/consultancy project you must complete 'An application for approval of an amendment to an existing application'.

The approval of the proposed research/consultancy project applies to the following site.

Project site: **University of East London**

Principal Investigator / Local Collaborator: Mr Samuel Kwakye

Approval is given on the understanding that the UEL Code of Practice for Research and the Code of Practice for Research Ethics is adhered to.

Any adverse events or reactions that occur in connection with this research/consultancy project should be reported using the University’s form for Reporting an Adverse/Serious Adverse Event/Reaction.

The University will periodically audit a random sample of approved applications for ethical approval, to ensure that the projects are conducted in compliance with the consent given by the Ethics and Integrity Sub-Committee and to the highest standards of rigour and integrity.

To support your research, UEL's Public and Community Engagement (PCE) function can provide access to external organisations, collaborative opportunities, support for participatory research, dissemination of your findings and policy development, throughout the life cycle of your project.

To discuss your research and explore what's possible, including identifying and support, please email: Natalie Freeman, Public and Community Engagement n.freeman@uel.ac.uk.

Please note, it is your responsibility to retain this letter for your records.

With the Committee's best wishes for the success of the project.

For further guidance and resources please check our Research Ethics Handbook.

Yours sincerely,

Fernanda Da Silva Hendriks

Research Ethics Support Officer

Ethics ETH2425-0284: Mr. Samuel Kwakye (High risk)

Expanded Stakeholder Responses by Thematic Category

The table below presents illustrative stakeholder responses from a diverse sample of stakeholders, grouped according to thematic categories (n ≈ 20):

Theme	Illustrative Responses
Visualization Clarity and Simplicity	“I prefer summaries instead of full plots.” “There are too many charts. I wasn’t sure which one to focus on first.” “Too many overlapping visuals. Simplicity would help.” “Vocabulary like ‘counterfactual’ doesn’t help, it confuses.” “The score is there, but I need to understand what it’s telling me.” “Give me visuals that are quick to read.” “I want simple bar graphs, not complicated dashboards.” “Avoid technical colours or gradients and keep it simple.”
Interpretability and Cognitive Load	“Sometimes it felt more confusing than helpful.” “It wasn't clear what I was meant to look at first.” “I’ve been told what the model predicts, but not why it matters.” “It made me question whether I was interpreting things right.” “The dashboard is dense. I worry teachers will click and guess.” “Sometimes it feels like stats for stats’ sake. I want insights I can act on.” “I was overwhelmed by how many tabs I had to go through.”
Subject-Specific Needs	“Maths and reading need different support plans. A one-size-fits-all explanation doesn’t work.” “Reading-level predictions need more context.” “What helps me in English isn’t the same as in Maths.” “In Maths, I want to see trends in number-based skills.” “Group these by subject so it’s clearer.” “tell me what this means for numeracy.

	What's affecting it?" "If the model suggests a low score in reading, I want it linked to literacy strategies."
Fairness, Labelling and Ethics	"Even if they've got the same grade, the justification might differ, that matters." "We shouldn't reinforce disadvantages with data." "I'm worried about reinforcing labels. Deficit thinking is a real risk." "It's affecting how I think about kids, that's not always good." "Sensitive categories shouldn't be so prominent." "Demographics alone don't tell the full story." "Why is EAL even in the chart? That feels like stereotyping." "Feels like some students are being profiled."
Practical Usefulness and Actionability	"If attendance is low and predicted grades are down, then what?" "Show historical context, how is this different now?" "If I can act on it, it's useful, otherwise not useful." "The predicted score helps, but I want to know what to do next." "Useful if it helps plan interventions earlier." "Highlight the top 2 things I can do per student." "Show me charts that matter for decision-making, not just data."