Machine Learning-Based Techniques for Assessing Critical Factors for European Tick Abundance

Abin Zorto[1], Samantha Lansdell[2], Misaki Seto[2], Edessa Negera[2], Mhd Saeed Sharif[1,*], Sally Cutler[2]
[1]Computer Science and DT, ACE, University of East London, University Way, London, UK
[2]School of Health, Sport & Bioscience, University of East London, London, UK
Email: u2091940@uel.ac.uk (A.Z); slansdell@uel.ac.uk (S.L); u1725457@uel.ac.uk (M.S);
e.n.gobena@uel.ac.uk(E.G); s.sharif@uel.ac.uk(S.S); s.cutler@uel.ac.uk (S.C);
*Corresponding author

***Abstract*—Tick-borne diseases are a significant health risk to humans and animals worldwide. It is important to understand the environmental and climatic factors that contribute to tick occurrence rates in order to reduce the proliferation of tick borne diseases. Using machine learning and spatial indexing techniques, this study covers tick occurrence rates in Europe over the last 20 years to understand the environmental and climatic factors that contribute to *Ixodes ricinus* tick abundance. We used biodiversity databases to study land cover categories, climate, vegetation index, and sociological factors. Areas with agriculture and natural vegetation, especially broad-leaved forests, had the strongest tick correlation. Waterways and pastures also showed significant positive correlations, indicating tick habitats. Ticks have moderate associations with urban green spaces, industrial units, and mixed forests suggesting their presence in ecologically disturbed habitats. Geoclimatic factors namely Normalised Difference Vegetation Index and rainfall, showed weak to negative correlations with tick population, indicating that they were less important than previously assumed. Linear Regression, Decision Tree, Random Forest, and Support Vector Machine were compared. We found that feature set and outlier presence significantly affected model performance. After removing outliers, Linear Regression performed best for land use features, with a R^2 value of 0.81, NRMSE of 1.56, SI of 1.56, and MAPE of 1.22. Outlier exclusion improved the model performance results. This research emphasises the importance of specific land uses in predicting the dynamics of tick population. Our findings lay the groundwork for focused intervention strategies to reduce the spread of tick-borne diseases using an innovative and intelligent approach, while also emphasising the need for further investigation into the complex interactions between environmental factors and tick abundance.**

***Index Terms*—Tick Borne Diseases, *Ixodes ricinus*, Linear Regression, Decision Tree, Random Forest, Support Vector Machine**

I. Introduction

Ticks have a vast influence on global human and veterinary health [1]. The survival of these parasites depends on them feeding from a suitable host to obtain their blood meals. During this process of feeding, they can transmit several pathogens. In Europe, the most frequently transmitted tick-borne pathogen is *Borrelia burgdorferi* sensu lato (the causative agent of Lyme Disease/Borreliosis). Other pathogens of increasing concern in Europe include Tick-borne Encephalitis virus, *Anaplasma phagocytophilum*, *Rickettsia* species and *Babesia* species [2, 3, 4].

In Europe, *Ixodes ricinus* (the most common tick species) populations are increasing over time thus raising the risk of pathogen transmission [4, 5]. It has been hypothesised that climatic factors like temperature and rainfall and habitat factors including vegetation, land cover features, and habitat fragmentation might play a major role as well [5]. However, to fully comprehend this, more insights are needed.

Machine learning applications have become increasingly valuable in analysing vector-borne infections because they can handle complex multidimensional datasets [6]. These tools will allow for better prediction and early warning systems for disease outbreaks since they analyse data from multiple sources hence promoting proactive public health responses [7]. Moreover, machine learning techniques help capture important environmental, climatic, and socio-economic predictors which govern the spread of vector-borne disease, leading to an understanding of their specific impacts [7]. Additionally ML models can predict vector population dynamics and abundance necessary for assessing disease transmission potential. They also do classification as well as diagnosing vector-borne diseases based on symptoms lab results among other clinical data thereby improving diagnosis speed and accuracy [8]. Furthermore, machine learning tools offer highly detailed, spatially explicit risk maps that optimise other public health interventions such as vector control by predicting their potential impact [9], enabling understanding of where diseases are likely to occur in space and time. Machine learning also integrates different types of data such as climate data, satellite imagery and social media thereby enhancing its predictive power and enabling real-time surveillance systems for disease activity that can constantly monitor and predict [7, 9]. In areas with limited surveillance data, sparse datasets can yield valuable insights through machine learning approaches that address data limitations [6]. Overall, applying ML in analysing vector-borne diseases brings significant progress to public health decision-making and disease management.

A. Machine Learning algorithms for the analysis of vector borne infection

A study used Bayesian priors and a linear model applied to spatially explicit classification models that related the occurrence probability of *I. ricinus* ticks to environmental factors [10].

Their method, which involved blanket-dragging across different sites in the Netherlands for tick collection, facilitated effective mapping of environmental risks associated with tick presence.

In another unique study, random forests were linked with Poisson regression models to analyse volunteered tick bite reports from the Netherlands [11]. This technique effectively addressed issues like zero-inflation and over-dispersion that resulted in accurate spatial prediction of risk areas inhabited by ticks. However, as this study focuses on tick bites (relying upon tick-host interaction) rather than just tick presence, it is arguably not a true indicator of tick abundance in a particular region.

B. Machine learning for tick occurrence rates analysis

In a study carried out in Scandinavia, Boosted Regression Tree (BRT) modelling was used to predict *I. ricinus* abundance. Ticks were counted at ten sites between August-September 2016, a time when larval ticks would have predominated, yet the researchers chose to focus upon nymph abundance to overcome the patchy distribution of larval ticks. This data from each location was combined with temperature, rainfall and land cover category [12]. However, the limitation of this study is the fact that it only considered two years. Furthermore, there was no attempt to incorporate spatial autocorrelation, which might undermine the accuracy of the model. The scope of their research was also limited because they employed only boosted regression trees for their machine learning analysis.

MaxEnt which is a tool for species distribution and environmental niche modeling [13] was used in Italy over two years to determine areas suitable for *I. ricinus* ticks with respect to multiple predictors such as temperature, rainfall and vegetation index. However, the most significant predictor variables were vegetation index and temperature [14]. The study, however, had serious methodical limitations. The modeling was based exclusively on presence data for MaxEnt, which is a major source of potential bias [15]. This is a typical weakness in MaxEnt since it only requires presence and background data and does not require any information on true absences. Also, the validation of the model was quite limited since it was done only in 2017 on 10 new sites without including the spatial autocorrelation effects within the dataset. The relationship between tick abundance and habitat fragmentation has also been previously investigated. A Spanish study concluded through statistical modelling that tick population growth is associated with habitat connectivity (low fragmentation) [16].

A 2019 study conducted by Garcia-Marti et al. [11], have implemented Random Forest (RF) together with 4 Poisson-family count data models to classify data into homogeneous segments. This fusion of models was able to better the standard RF model for highly skewed and excessive zero- count data. However, the count data models did not reach convergence in 5-9% of the leaf nodes due to high data sparsity. Details of the study, for instance the 60-40 split of data into training and testing with no cross validation might have also affected the way the model might be generalised in real world conditions [11].

Lihou and Wall [17] utilised random forest models to analyse machine learning data, primarily derived from retrospective questionnaire responses from farmers. This methodology has certain drawbacks. Bias might result from the farmers' ability to remember and identify ticks, and not from conducting objective scientific sampling, resulting in recall and reporting biases affecting data quality. Moreover, the analysis of the data did not generate predictions with respect to spatial autocorrelation, which may influence prediction 'accuracy' within the broader geographic scope of the study. The sample, however, was quite small compared to the high number of predictions mapped out across Great Britain covering 926 farms. The temporal scope is also limited with one main data set at hand for one year (2017- 2018) hence limits understanding of long-term trends.

C. Challenges faced in predicting tick occurrence rates

It is therefore clear that there are several challenges that are faced when attempting to utilise tick occurrence data for analysis among them the use of incomplete data, biased data and also using non standardised data which makes prediction difficult. The majority of studies rely on different sources like reviews from literature, reports given by health departments or even personal communications rather than employing systematic sampling methods that are uniform across all areas. Additionally, there is no comprehensive county-level data concerning metrics such as density of infected host-seeking nymphal ticks coupled with temporal uncertainties associated with historical establishment of these tick populations thereby rendering current databases unreliable [18].

Furthermore, model generalisation poses a challenge as models are often developed for particular geographic locations or time periods that fail when applied in new regions or under different temporal conditions [19]. This therefore calls for advanced techniques like careful feature selection and cross-validation to enable models to extrapolate well on unseen data. Tick occurrences are driven by ecological dynamics in a complex manner which involves complex climate-land use-host population-environmental factors connections. These relationships not only vary spatially and temporally but also make it difficult to prioritise which factors are more important than the others [20]. In addition, difficulties related to scale and resolution of analysis further complicate the modeling process since tick population dynamics occur at multiple scales and coarse national data may not adequately capture fine-scale patterns [21]. Indeed, the NUTS3 resolution which is a hierarchical system for dividing up the economic territory of the European Union for statistical purposes, typically with populations between 150,000 and 800,000 may not suffice to capture the focal ecological niches that help ticks to thrive [22]. Furthermore, climate change adds another layer of complexity by potentially disrupting past associations between environmental variables and tick occurrences [20].

D. Limitations of current research

In this study we attempt to resolve some of the shortcomings of current studies with respect to tick abundance.

1) **Time period:** In this study the ticks have been collected

over a period of 20 years. In comparison, current research tends to look at a period of 1-2 years which may not be enough for studying trends in tick behaviour over long periods of time.

2) **Spatial Autocorrelation:** Various studies which considered occurrence rates of ticks failed to incorporate spatial autocorrelation in the analysis of the observed tick behaviour. In our research, we employ algorithmic clustering techniques based purely on the geographical coordinates of each tick occurrence.
3) **Data Quality:** Some past studies have used data that is vulnerable to bias like farmer questionnaires. Some standardisation procedures and approaches are not adhered to while carrying out research processes hence compromising the reliability of the resulting data.
4) **Model Comparison:** A common limitation in many of these studies is the use of one machine learning approach without any attempt to compare the different approaches. We evaluate different algorithms (Linear Regression, Decision Tree, Random Forest and Support Vector Machine) in a bid to find out the most appropriate one to use.
5) **Comprehensive Feature Set:** Inherent environmental aspects like land use type have been included alongside geoclimatic and other observational aspects. This provides a better characterisation of the training dataset in comparison with current studies where only one subset of features is used
6) **Systematic Verification:** In contrast to current research this study employs validation techniques in a bid to achieve robustness of results.
7) **Broader Scope of Performance Assessment:** Multiple metrics (NRMSE, SI, MAPE, R^2) designed to assess certain characteristics, are utilised with the aim of enhancing the trustworthiness of model results.

II. Methodology

Briefly, our methodology encompasses data collection, preprocessing, feature selection, and the application of various machine learning techniques.

Data Description

A total of 5590 individual tick occurrence records were obtained from three online databases: National Biodiversity Network Atlas, Global Biodiversity Information Facility and Vectormap. Other records were obtained directly from Institute of Public Health, Albania.

Records were exported if they met the following criteria: *I. ricinus* species, 2000-2019 and European location.

The dataset was filtered to include the following information for each record: occurrence ID, date of occurrence, data source, latitude and longitude coordinates. For each record, data relating to further variables were added (specific to location coordinates and date of occurrence). Firstly, temperature and rainfall climatic variables were added from an online Weather Data and API resource (VisualCrossing). In relation to habitat, Normalised Difference Vegetation Index (NDVI) value and Land Cover category were added from an online repository (EcoDataCube). In the end, from the EcoDataCube repository, a Discontinuous Urban Fabric (%) value was added as a measure of physical barriers between habitats – thus representing habitat fragmentation.

B. Data Preprocessing

Initial cleaning of the data by removing entries that lacked geographical coordinates, NDVI, or land use variables was undertaken. For temporal feature engineering, we simplified date information to capture the day of the year, enabling the identification of seasonal variations. This conversion of date features into a single 'day of year' feature was justified by its ability to capture cyclical patterns within a year while reducing dimensionality. We standardised missing or inconsistent dates to a common placeholder before converting them into a uniform format. In the feature selection phase, we chose relevant variables for analysis, including geographic coordinates, environmental measures, and temporal indicators. Categorical data, such as land use methods, were transformed into binary format for easier processing. For outlier removal, we implemented a 95th percentile threshold instead of the more common 75th. This decision was justified by the nature of our data in which potentially important data points can be retained while still eliminating the most extreme outliers. After outlier removal the number of clusters reduced from 343 to 333 while the total tick frequency decreases from 24,346 instances of tick occurrences to 4928 instances. Finally, we applied the 10 folds cross-validation aimed at improving the robustness of our models. By applying this strategy, the assessment and reporting of modeling performance is done with greater confidence and risk of modeling overfitting is reduced.

C. Methodology Workflow

Fig. 1 was created to illustrate the Machine learning workflow for tick occurrence prediction. The process started with cleaning and selecting data so that only high-quality data will be used for training models. Then data clustering is performed by grouping similar data points together to find patterns and increase model accuracy.

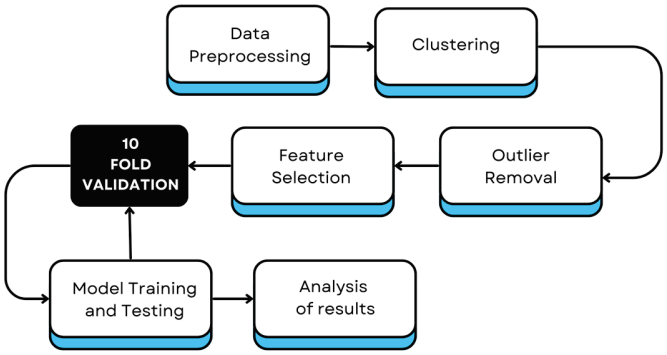

Fig. 1. Machine learning workflow for tick occurrence prediction

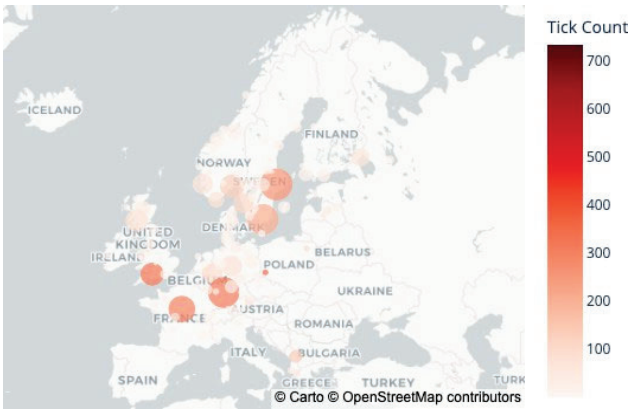

Fig. 2. Tick occurrence Bubble Map detailing DBSCAN clusters

After that, the dataset is split into two parts: 80% for training and 20% for testing. Different machine learning models are trained with the training set until all the models have been fitted and evaluated against some predefined performance metrics where the best performing model is selected as per those criteria. Finally, the selected model undergoes final fine-tuning through extra optimisation parameter changes in order to achieve higher precision rates during predictions. This well-structured approach ensures sound development of models which help in identifying key predictors for tick abundance therefore leading targeted interventions towards tick borne diseases management.

D. Clustering and Feature Selection

Three geospatial clustering algorithms were used to identify areas influencing tick collection: K-Means DBSCAN Agglomerative Hierarchical Clustering (AHC). These methods helped create distinct neighborhoods which enable localised analysis of environmental variables. This paper focuses on DBSCAN clustering selected for its ability to identify clusters of arbitrary shape without requiring a predefined number of clusters. In our research, we set epsilon (ϵ) equal to 0.005 degrees (approximately 500 meters at equator). A smaller ϵ could detect clusters of ticks close to each other indicating specific areas or host availability zones while minimum samples =1 means any single point forms its own cluster resulting frm the absence of other points around it meeting required distance threshold. Increasing minimum samples would ensure significant clusters but if too many points were not within range of other points, it may fail to indicate accurate locations of tick occurrence.

There were 344 clusters produced by the algorithm (Fig.2). All these clusters were analysed to establish the most important factors contributing towards tick occurrence rates. In this study, machine learning models were used to investigate the correlation between different environmental variables and tick occurrences in a methodologically rigorous way. The analysis was divided into three groups of features:

a) Observational factors: These include direct observations and counts of ticks which are important for understanding immediate impacts on environment but can be over-simplistic when it comes to predictive modelling because they correlate directly with presence/absence data.

b) Geo-Climatic factors: Geographical (latitude and longitude) and climatic factors (temperature, rainfall, NDVI) relating to environment of tick habitats are combined and termed as Geo-Climatic Factors.

c) Land use factors: Land Use Factors include agriculture, forest cover, and urban areas which give an idea how the man-made environments such as modifications disturbs the normal pattern of ticks.

E. Correlation Analysis

Correlation analysis was done to measure strength and direction between tick count with different environmental features. This helped in identifying variables that are highly associated with occurrence of ticks hence guiding selection process for feature engineering prior machine learning model building.

F. Evaluation Metrics

In this study, we utilised four different evaluation metrics.

1) The Normalised Root Mean Square Error (NRMSE) which is a standardised variation of RMSE, making it usable across datasets of different scales.
2) The coefficient of determination (R^2) which illustrates the model's ability to understand the variance in the dataset.
3) The Scatter Index (SI) which is a dimensionless error measure calculated based on the mean of the observed values.
4) The Mean Absolute Percentage Error (MAPE) which is a metric used to measure the percentage error of a model's predictions, providing insight into how inaccurate the model might be.

G. Algorithm selection

Linear Regression was selected because of its transparency and the capability to fit the linear relationship among different variables. It acts as a standard on which other more machine capability models can be assessed and in addition provides information on value of attributes.

The Decision Tree Regressor captures non-linear relationships and interactions among features.

Decision trees overfit by default but Random Forests aim to reduce this issue through ensemble learning thereby increasing generalisation error. This algorithm builds multiple decision trees then takes vote from each tree to give final prediction for regression setting.

SVM Regressor works well in high-dimensional spaces where there may be many independent variables with complex relationships that cannot be modeled using linear functions alone.

III. Results

Initially, we examined the correlations between different environmental factors and tick abundance, then compared the performance of our machine learning models across various feature sets and data preprocessing scenarios.

A. Model Performance Across Feature Sets

Table I presents the performance metrics (MAPE, NRMSE, SI, and R^2) for all models across the different feature sets, both with and without outliers.

1) Geoclimatic Features: With outliers present, models trained on geoclimatic features showed poor performance. Linear Regression had a high MAPE of 57.73 and a negative R^2 of -0.07. Decision Tree and Random Forest models performed similarly poorly (MAPE = 39.37 and 52.24, R^2 = -1.03 and -0.32, respectively). The SVM model showed unusually low MAPE (0.78) but still a negative R^2 (-0.02), suggesting potential issues with the model's fit.

Removing outliers led to some improvements, particularly for Linear Regression and Random Forest (MAPE reduced to 8.55 and 6.73, respectively). However, R^2 values remained negative or close to zero for all models, indicating that geoclimatic features alone were poor predictors of tick presence.

2) Observational Features: The presence of observational features in our dataset led to significant overfitting, particularly evident in the Linear Regression (LR) model's perfect fit (R^2 = 1.00) when outliers were included. This overfitting stems from the direct one-to-one relationship between tick occurrences and observational instances. Each recorded tick presence corresponds precisely to one observation event, creating an artificial simplification of the prediction task. Consequently, models, especially LR, achieve deceptively high-performance metrics by essentially memorising this direct relationship rather than learning generalisable patterns. The near-perfect scores (R^2 = 0.96) for Decision Tree (DT) and Random Forest (RF) models after outlier removal further illustrate this issue.

3) Land Use Cover Features: Models trained on land use cover features showed better performance compared to geoclimatic features. With outliers, Linear Regression performed reasonably well (MAPE = 3.17, R^2 = -6.06), while Decision Tree and Random Forest showed similar performance (MAPE around 3, R^2 around 0.15). SVM again showed a low MAPE (0.76) but a near-zero R^2.

Outlier removal improved performance across all models for land use features. Linear Regression showed significant improvement (MAPE = 1.22, R^2 = 0.81), as did Decision Tree and Random Forest (MAPE around 1.7, R^2 around 0.40). SVM also improved but remained the weakest performer (MAPE = 0.49, R^2 = 0.05).

4) Excluding Observation Features: The results for models trained without observation features were identical to those trained on land use cover features, both with and without outliers. This suggests that the land use cover features were the primary drivers of model performance when observation features were excluded. The exclusion of observational features in the models focuses the analysis on environmental and geoclimatic factors, revealing that land use features almost exclusively influence tick occurrences.

IV. Discussion

The results of our investigation provide several key observations regarding the reasons for the variation in tick populations with respect to Europe.

A. Importance of feature sets

Land use factors such as land principally occupied by agriculture, broad-leaved forests and water bodies, resulted in the greatest relationship which indicates such areas are essential for the ticks Fig. 3. On the other hand, geoclimatic factors which include temperature and rainfall were surprisingly very weakly to slightly negatively correlated related.

This correlation analysis is crucial in isolating those environmental and climate factors that are most likely to be associated with tick occurrences. Positive high correlations with particular land use types further identifies critical land areas that have high concentrations of ticks informative to targeted control measures. For example, there is need to focus on agricultural lands, particularly that host natural vegetation and broad leaved forests, as they tend to be more conducive for supporting tick populations.

On the other hand, explanation of reasons for some features being associated with low correlation or negative correlation is necessary in improving the models and their predictions and descriptions. To give an example, the rainfall in general had negative relationship, and this could be particularly useful

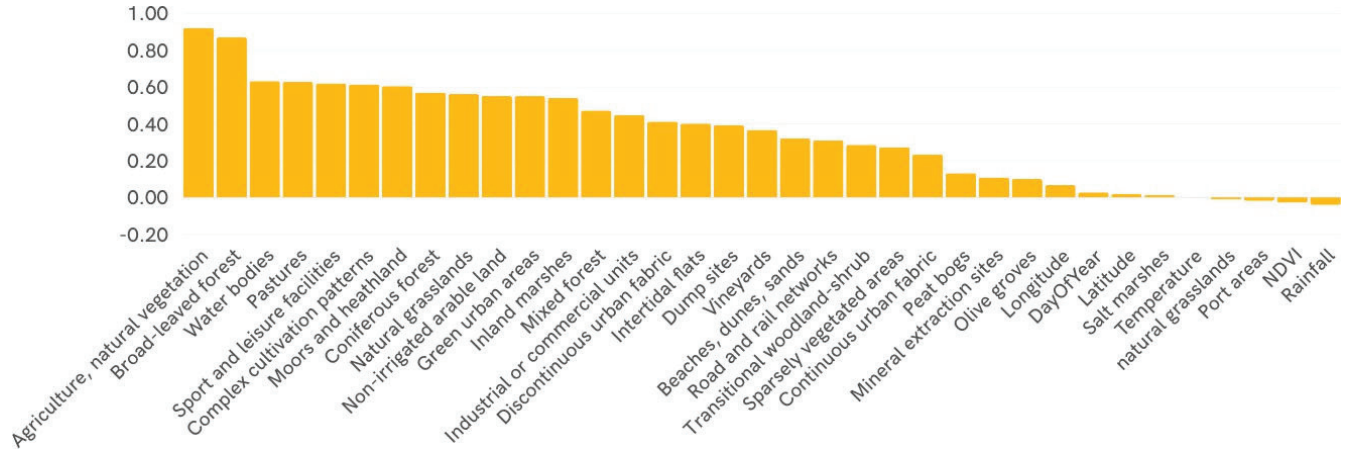

Fig. 3. Correlation Bar Chart of Geo-climatic and Land Use Features

TABLE I. DBSCAN Results for Selected Models

		Outliers DBSCAN				No Outliers DBSCAN			
Feature Set	Model	MAPE	NRMSE	SI	R2	MAPE	NRMSE	SI	R2
Geoclimatic	LR	57.73	7.56	7.56	-0.07	8.55	3.72	3.72	-0.11
	DT	39.37	10.44	10.44	-1.03	8.73	5.23	5.23	-1.19
	RF	52.24	8.43	8.42	-0.32	6.73	3.67	3.67	-0.08
	SVM	0.78	7.39	7.32	-0.02	0.74	3.63	3.53	-0.06
Land use	LR	3.17	19.46	19.36	-6.06	1.22	1.56	1.56	0.81
	DT	2.99	6.70	6.67	0.16	1.73	2.78	2.76	0.38
	RF	2.86	6.78	6.75	0.14	1.71	2.70	2.69	0.42
	SVM	0.76	7.37	7.31	-0.01	0.49	3.45	3.37	0.05
Observation	LR	0.00	0.00	0.00	1.00	0.84	2.07	2.06	0.66
	DT	0.04	6.38	6.36	0.24	0.04	0.70	0.70	0.96
	RF	0.04	6.47	6.45	0.22	0.04	0.74	0.74	0.96
	SVM	0.73	7.38	7.32	-0.02	0.63	3.50	3.40	0.02
Observation Excluded	LR	3.17	19.46	19.36	-6.06	1.22	1.56	1.56	0.81
	DT	2.99	6.70	6.67	0.16	1.73	2.78	2.76	0.38
	RF	2.86	6.78	6.75	0.14	1.71	2.70	2.69	0.42
	SVM	0.76	7.37	7.31	-0.01	0.49	3.45	3.37	0.05

to explore reasons as to how precipitation influences tick dynamics and survival rates.

Analysis of random forest feature importance reveals key factors influencing tick abundance across geoclimatic and land use categories Table II. Location, for instance latitude is among the most important geoclimatic feature (0.33) followed by the rainfall whose weight was also significant (0.26). The day of year (0.14) NDVI (0.1) and temperature (0.09) do contribute, though 'slightly' with longitude (0.08). In terms of land use, areas principally occupied by agriculture show the highest importance (0.52) and in addition broad-leaved forests have been found to be another major factor (0.29). Water bodies (0.09), complex cultivation patterns (0.03), other such agricultural activities entailed exert moderate effect, but others such as land cover types including urban areas, natural grasslands, pastures showed only miniscule but measurable effects on tick model output.

B. The Impact of Feature Sets and Outliers on model performance

The presence of outliers always reduced the efficiency of the geoclimatic features models. In the most models after outliers' removal, the situation was much better than before, especially for the Linear Regression model, with the R^2 rising from -0.07 to -0.11 and NRMSE declining from 7.56 to 3.72.

Land use features had greatest influence with outlier removal. For instance, in the case of LR model, removal of outliers resulted in a fundamental change with R^2 moving from
-6.06 to 0.81 while NRMSE reduced from 19.46 to 1.56.

Observation features showed the most striking impact. With outliers present, the LR model attained a perfect fit (R^2 =1.00 and NRMSE=0.00), a sign of gross overfitting, where the model conforms to noise rather than to underlying data structure.

When the observation features were detached from the dataset, the results corresponded with the results obtained with land use features and reiterated the need to incorporate land use data.

TABLE II. MOST SIGNIFICANT FEATURES BY CATEGORY (NO OUTLIERS)

Observation	Imp.	Geoclimatic	Imp.	Land Use	Imp.
DCM_HUMAN OBSERVATION	0.66	Latitude	0.33	Land princ. occ. by agriculture	0.52
DCM_ RESEARCH STUDY	0.26	Rainfall	0.26	Broad-leaved forest	0.25
DCM_ PRESERVED SPECIMEN	0.07	DayOfYear	0.14	Water bodies	0.09
DCM_ MATERIAL SAMPLE	0.00	NDVI	0.10	Complex cultivation patterns	0.03
DCM_MACHINE OBSERVATION	0.00	Temperature	0.09	Green urban areas	0.03
DCM_LIVING SPECIMEN	0.00	Longitude	0.08	Natural grasslands	0.02
				Moors and heathland	0.02
				Pastures	0.02
				Sport and leisure facilities	0.02
				Inland marshes	0.01

C. Model Performance Across Feature Sets

It was determined that even within the data typically used for the construction of regression models with a complex structure (for example DT, RF or SVM), the Linear Regression model (LR) was the best fit in most scenarios, especially after outlier exclusion. This is surprising in some regard and may be explained by the mainly linear link between selected features and tick abundance, which we made more efficient than necessary through preliminary feature selection.

While we did take into consideration cross-validation techniques to limit any potential effects of overfitting, it may be prudent to apply external datasets for further validation.

There are some contradictions in our analysis of land use and climate factors with some previous studies. We found consistent and powerful relationships between land use and its features and much weaker climatic influences, while Boulanger et al. [23] highlighted temperature as the most relevant driver. This discrepancy may be due to the scale of our Europe-wide, 20-year study capturing broader patterns that may overshadow local climatic influences, as well as the complex interaction effects between land use and climatic factors.

More investigation is required to understand the relationship between rainfall and tick abundance. Studies show Nymphal and adult ticks were more abundant when there had been high cumulative rainfall in the prior months. However, larval abundance did not appear to be sensitive to prior rainfall, suggesting a complex, non-linear response to different rainfall patterns [24].

V. CONCLUSION AND FUTURE WORK

Based on these results, it is evident that adequate predictive models can only be built if sufficient preprocessing measures such as outlier detection and exclusion are undertaken. This assertion is particularly true, in Linear Regression models, where the performance of the model improves tremendously after outlier removal. We demonstrate that it is essential to use sophisticated machine learning methods for predictive modeling in environmental and public health settings. After outlier removal, the Linear Regression model performs well across different feature sets, particularly with land use features (R^2 of 0.81, NRMSE of 1.56). This study emphasises the role of specific land use types on tick population dynamics and paves the way towards the development of more efficient tick-borne disease control measures.

A. Future Work

1) Model architecture: The variability in model performance across different setups and conditions indicates the necessity for dynamic modeling techniques that adapt to specific dataset characteristics. Future work may look into the hybridisation or ensembling of different techniques to enhance the level of prediction accuracy and robustness.

2) Other Outlier Detection Techniques: Here, we used a modified IQR method to detect outliers, however there are a number of other techniques that can also be suggested for use in future research. Future work might include, the Z score method, the Local Outlier Factor (LOF) method, and Isolation Forest to name a few. These methods however lie at the contradictory poles of outlier detection which would form a very relevant outlier detection baseline in subsequent analyses.

3) Real Time Data: Future studies can also aim at empowering the studies with real time data. This approach will not only increase prediction accuracy but also strengthen public health interventions against tick-borne diseases. Continuous refinement of these models is necessitated by ongoing environmental changes, with a view to enhancing prediction generalisability and adapting to evolving ecological conditions that influence tick populations.

REFERENCES

[1] B. Cull, M. E. Pietzsch, K. M. Hansford, E. L. Gillingham, and J.

M. Medlock, “Surveillance of British ticks: An overview of species records, host associations, and new records of *Ixodes ricinus* distribution,” *Ticks Tick Borne Dis.*, vol. 9, no. 3, pp. 605–614, Mar. 2018.

[2] J. Brites-Neto, K. M. R. Duarte, and T. F. Martins, “Tick-borne infections in human and animal population worldwide,” *Vet World*, vol. 8, no. 3, pp. 301–315, Mar. 2015.

[3] C. F. Köhler, M. L. Holding, H. Sprong, P. A. Jansen, and J. Esser, “Biodiversity in the Lyme-light: ecological restoration and tick-borne diseases in Europe,” *Trends Parasitol.*, vol. 39, no. 5, pp. 373–385, May 2023.

[4] J. Gray, O. Kahl, and A. Zintl, “What do we still need to know about *Ixodes ricinus*?” *Ticks Tick Borne Dis.*, vol. 12, no. 3, p. 101682, May 2021

[5] J. M. Medlock, K . M. Hansford, A . Bormane, M. Derdakova, A . Estrada-Peña, J . -C. George, I. Golovljova, T. G. T. Jaenson, J.-K. Jensen, P. M. Jensen, M. Kazimirova, J. A. Oteo, A. Papa, K. Pfister, O. Plantard, S. E. Randolph, A. Rizzoli, M. M. Santos-Silva, H. Sprong, L. Vial, G. Hendrickx, H. Zeller, and W. Van Bortel, “Driving forces for changes in geographical distribution of *Ixodes ricinus* ticks in Europe,” *Parasit. Vectors*, vol. 6, p. 1, Jan. 2013.

[6] S. Raizada, S. Mala, and A. Shankar, Vector-Borne Disease Outbreak Prediction Using Machine Learning Techniques. Cham: Springer International Publishing, 2021, pp. 227–241. [Online]. Available: https://doi.org/10.1007/978-3-030-66519-7 9

[7] D. P. C. Peters, D. S. McVey, E. H. Elias, A. M. Pelzel-McCluskey, J. D. Derner, N. D. Burruss, T. S. Schrader, J. Yao, S. J. Pauszek, J. Lombard, and L. L. Rodriguez, “Big data–model integration and AI for vector-borne disease prediction,” *Ecosphere*, vol. 11, no. 6, Jun. 2020.

[8] S. G. Shaikh, B. S. Kumar, G. Narang, and N. N. Pachpor, “Hybrid machine learning method for classification and recommendation of vector-borne disease,” *Journal of Autonomous Intelligence*, vol. 7, no. 2, Dec. 2023.

[9] O. E. Santangelo, V. Gentile, S. Pizzo, D. Giordano, and F. Cedrone, “Machine learning and prediction of infectious diseases: A systematic review,” *Machine Learning and Knowledge Extraction*, vol. 5, no. 1, pp. 175–198, Feb. 2023.

[10] A. Swart, A. Ibañez-Justicia, J. Buijs, S. E. van Wieren, T. R. Hofmeester, H. Sprong, and K. Takumi, “Predicting tick presence by environmental risk mapping,” *Front Public Health*, vol. 2, p. 238, Nov. 2014.

[11] I. Garcia-Marti, R. Zurita-Milla, and A. Swart, “Modelling tick bite risk by combining random forests and count data regression models,” *PLoS One*, vol. 14, no. 12, p. e0216511, Dec. 2019.

[12] L. Jung Kjær, A. Soleng, K. S. Edgar, H. E. H. Lindstedt, K. M. Paulsen, Å. K. Andreassen, L. Korslund,V. Kjelland, A. Slettan, S. Stuen, P. Kjellander,M. Christensson, M. Teräväinen, A. Baum, K. Klitgaard, and R. Bødker, “Predicting the spatial abundance of *Ixodes ricinus* ticks in southern Scandinavia using environmental and climatic data,” *Sci. Rep.*, vol. 9, no. 1, p. 18144, Dec. 2019.

[13] C. Merow, M. J. Smith, and J. A. Silander Jr, “A practical guide to maxent for modeling species’ distributions: what it does, and why inputs and settings matter,” *Ecography*, vol. 36, no. 10, pp. 1058–1069, 2013.

[14] M. Signorini, A.-S. Stensgaard, M. Drigo, G. Simonato, Marcer, F. Montarsi, M. Martini, and R. Cassini, “Towards improved, cost-effective surveillance of *Ixodes ricinus* ticks and associated pathogens using species distribution modelling,” *Geospat. Health*, vol. 14, no. 1, May 2019.

[15] Y. Fourcade, J. O. Engler, D. R¨odder, and J. Secondi, “Mapping species distributions with maxent using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias,” PloS one, vol. 9, no. 5, p. e97122, 2014.

[16] A. Estrada-Peña, “The relationships between habitat topology, critical scales of connectivity and tick abundance *Ixodes ricinus* in a heterogeneous landscape in northern Spain,” *Ecography*, vol. 26, no. 5, pp. 661–671, Oct. 2003.

[17] K. Lihou and R. Wall, “Predicting the current and future risk of ticks on livestock farms in Britain using random forest models,” *Vet. Parasitol.*, vol. 311, p. 109806, Nov. 2022.

[18] K. J. Kugeler and R. J. Eisen, “Challenges in predicting Lyme disease risk,” *JAMA Netw Open*, vol. 3, no. 3, p. e200328, Mar. 2020.

[19] H. S. Tiffin, E. G. Rajotte, J. M. Sakamoto, and E. T. Machtinger, “Tick control in a connected world: Challenges, solutions, and public policy from a United States border perspective,” *Trop Med Infect Dis*, vol. 7, no. 11, Nov. 2022.

[20] O. Sparagano, G. Földvári, M. Derdáková, and M. Kazim´irova´, “New challenges posed by ticks and tick-borne diseases,” *Biologia*, vol. 77, no. 6, pp. 1497–1501, Jun. 2022.

[21] A. M. Gardner, N. C. Pawlikowski, S. A. Hamer, G. J. Hickling, J. R. Miller, A. M. Schotthoefer, J. I. Tsao, and B. F. Allan, “Landscape features predict the current and forecast the future geographic spread of Lyme disease,” *Proc. Biol. Sci.*, vol. 287, no. 1941, p. 20202278, Dec. 2020.

[22] P. D’Urso, L. D. Giovanni, F. G. Sica, and V. Vitale, “Measuring competitiveness at nuts3 level and territorial partitioning of the Italian provinces,” *Social Indicators Research*, vol. 173, no. 1, pp. 9–51, 2024.

[23] N. Boulanger, D. Aran, A. Maul, B. I. Camara, C. Barthel, M. Zaffino, M.-C. Lett, A. Schnitzler, and P. Bauda, “Multiple factors affecting *Ixodes ricinus* ticks and associated pathogens in European temperate ecosystems (northeastern France),” *Sci. Rep.*, vol. 14, no. 1, p. 9391, Apr. 2024.

[24] F. Keesing, R. S. Ostfeld, T. P. Young, and B. F. Allan, “Cattle and rainfall affect tick abundance in central Kenya,” Parasitology, vol. 145, no. 3, pp. 345–354, 2018.