ISSN: 2963-2536 (media online)
Proceedings of Malikussaleh International Conference on Multidisciplinary Studies (MICoMS)
Organized by center of Excellence of Technology Natural Polymer and Recycle Plastics

International Conference Proceedings 00007 (2024) DOI: https://doi.org/10.29103/micoms.v4.2024

Machine Learning Algorithms Comparison for Gender Identification

Aldo Januansyah. H[1], Muhammad Fikry*[2], Yesy Afrillia[3]

[1]Department of Informatic, Universitas Malikussaleh, Bukit Indah, Lhokseumawe, 24353, Indonesia, aldo.210170184@mhs.unimal.ac.id
[2]Department of Informatic, Universitas Malikussaleh, Bukit Indah, Lhokseumawe, 24353, Indonesia, muh.fikry@unimal.ac.id
[3]Department of Informatic, Universitas Malikussaleh, Bukit Indah, Lhokseumawe, 24353, Indonesia, yesy.afrillia@unimal.ac.id
*Correspondence: muh.fikry@unimal.ac.id;

Abstract: In this study, we presents a comprehensive analysis of gender identification methods utilising eight distinct classification models: K-Nearest Neighbors (KNN), Naive Bayes, Decision Tree, Random Forest, Logistic Regression, XGBoost, Support Vector Machine (SVM), and Neural Network. Gender identification is a critical task with significant applications in marketing, social analysis, and security systems, necessitating the exploration of various methodologies to achieve optimal performance. The dataset employed in this research underwent normalisation using the Min-Max scaling technique, which enhances the performance of classification models by ensuring that all features contribute equally, particularly when the data exhibits varying ranges of values. The results reveal that the K-Nearest Neighbors (KNN) model significantly outperformed the other models, achieving an impressive accuracy of 0.9758 with a support of 951, underscoring the effectiveness of the KNN algorithm in gender identification tasks and establishing it as a reliable choice for applications requiring high accuracy. Furthermore, the study emphasises the critical importance of selecting appropriate models in machine learning tasks and the substantial impact of data normalisation on model performance. Overall, this research provides valuable insights into the KNN algorithm, demonstrating its ease of implementation and exceptional effectiveness in achieving high precision in gender identification tasks, with implications for future research and practical applications across various fields

Keywords: classification models; data normalisation; gender identification; K-Nearest Neighbours; machine learning.

1. Introduction

Gender identification has become an essential task in various domains, including marketing, social analysis, security systems, and human-computer interaction. With the increasing availability of digital data, the demand for accurate and efficient gender classification methods has risen [1], enabling tailored services and enhanced user experiences. Traditional approaches to gender identification relied heavily on demographic and psychological surveys, which are often time-consuming and prone to bias. However, advancements in machine learning have

opened new avenues for automating this process with higher accuracy and efficiency, making it a focal point for research in recent years.

Machine learning algorithms offer promising solutions for gender identification by learning patterns from data and generalising these patterns to unseen cases [2]. Among the most widely used models are K-Nearest Neighbors (KNN), Naive Bayes, Decision Trees, Random Forest, Logistic Regression, Support Vector Machine (SVM), XGBoost, and Neural Networks. Each of these algorithms has its strengths and limitations, making it critical to assess their performance comprehensively. Previous studies have explored the use of these models in gender classification tasks, demonstrating varying degrees of success depending on the dataset characteristics and preprocessing techniques employed.

Data preprocessing, particularly feature scaling and normalisation, plays a crucial role in enhancing the performance of machine learning models. For instance, distance-based algorithms like KNN and SVM are highly sensitive to the range of input features, where differences in feature magnitudes can disproportionately influence model outcomes. The Min-Max scaling technique is widely adopted to address this issue by transforming feature values into a uniform range. The normalisation significantly boosts model performance by ensuring that all features contribute equally during the training process. This step is especially important when dealing with datasets containing features with varying scales.

Despite the significant progress made in gender identification using machine learning models, there remains a need for a thorough comparison of these algorithms to identify the most effective model for this task. Many studies have focused on individual models or small subsets of algorithms without providing a comprehensive evaluation across a broad range of classifiers. Additionally, the impact of data preprocessing techniques [3][4] such as Min-Max normalisation on different models' performance has not been fully explored [5]. Understanding these factors is vital to developing robust gender classification systems that can be applied across various contexts and applications.

This study aims to address these gaps by conducting a comprehensive analysis of eight popular machine learning models for gender identification. By applying Min-Max normalisation to the dataset and carefully evaluating each model's performance using standard metrics, this research provides valuable insights into the effectiveness of different classification algorithms. The findings of this study are expected to contribute to the growing body of knowledge on gender identification and offer practical guidance for selecting suitable machine learning models in various real-world applications.

2. Materials and Methods

This research aimed to analyse and compare the performance of eight distinct machine learning classification models for gender identification. The methodology was carefully designed and executed in several stages, including data preprocessing, model selection, training, evaluation, and performance metrics analysis. Each of these stages was integral to achieving a robust and accurate comparison of the chosen algorithms. The models selected for this study were K-Nearest Neighbors (KNN), Naive Bayes, Decision Tree, Random Forest, Logistic Regression, XGBoost, Support Vector Machine (SVM), and Neural Network, providing a diverse range of classifiers with varying underlying mechanisms.

2.1 Preprocessing Data

Data preprocessing is a crucial step in machine learning to enhance the quality of input data and ensure it aligns with the requirements of various algorithms. In this study, the dataset underwent several preprocessing steps, starting with data cleaning to handle missing or

inconsistent values. Following this, normalisation was performed using the Min-Max scaling technique [6][18]. This method transforms feature values into a fixed range, typically [0, 1], which is essential for distance-based algorithms like KNN and SVM [7]. By applying Min-Max normalisation, we ensured that features with larger ranges did not disproportionately affect the model, thereby improving the overall performance and convergence speed of the classifiers.

2.2 Feature Selection

Feature selection is a vital step in reducing model complexity and enhancing predictive performance by identifying the most relevant features. In this research, a statistical approach was employed to select key features that significantly contribute to gender classification [8]. Specifically, the mean values of the features were calculated, and an evaluation of their significance was conducted. The analysis identified three key features—meanfreq, meanfun, and IQR—as the most influential in distinguishing between male and female labels. This selection process effectively reduced the dimensionality of the dataset, enhancing the model's efficiency without compromising its accuracy.

2.3 Model Training and Evaluation

After data preprocessing and feature selection, the next phase involved training the chosen classification models. The dataset was split into training and testing. Each model was then trained on the training dataset and evaluated on the testing dataset to assess its performance. Hyperparameter tuning was performed for each model to optimise their configurations and improve accuracy. For instance, the number of neighbors in KNN, the maximum depth of Decision Trees, and the kernel type in SVM were fine-tuned to achieve the best results. The models' performance was evaluated using standard metrics, including accuracy, precision, recall, and F1-score, to provide a comprehensive comparison of their capabilities.

KNN is a distance-based algorithm that classifies data based on the majority label of the k nearest neighbors. The distance is cal culated using matrics such as Euclidean or Manhattan[9].

$$d(x, x_i) = \sqrt{\sum_{j}^{n} = 1((x_j - x_{ij})^2}$$

Naïve Bayes assumes that features are independent and calcutes probabilities using Bayes'theorem. It is suitable for large datasets with simple assumptions[10].

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

A Decision Tree builds a tree structure by splitting data based on metrics like Gini Impurity or Informtion Gain. Each branch represents a decision-making condition[11].

$$Gini = 1 - \sum_{i=1}^{c} p_i^2$$

Random Forest is an ensemble of multiple decision trees. The algorithm combines the predictions of individual trees to achieve higher accuracy and robustness[12][13].

$$\hat{f}(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x)$$

Logistic Regression predicts the probability of the target class by modeling a logistic relationship between independent variables and the dependent variable[14].

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

XGBoost is a boosting algorithm that iteratively improves model performance by adding decision trees and minimizing loss efficiently [15].

$$F_t(x) = F_{t-1}(x) + \eta h_t(x)$$

SVM finds the optimal hyperplane that separates data from different classes with maximum margin, ensuring better generalization[16].

$$min_{w,b} \frac{1}{2} \| w \|^2 \quad s.t \quad y_i(w . x_i + b) \geq 1$$

A Neural Network consists of layers of interconnected neurons. Each neuron processes information using an activation function to model complex relationships in data[17].

$$a^{(l)} = f(W^{(l)} a^{(l-1)} + b^{(l)})$$

Each model was trained using a consistent training procedure. The dataset was divided into training and testing subsets. In this study, the **random splitting technique** was applied, with a proportional split of 70:30, where 70% of the data was allocated for training and 30% for testing. The training subset was used to fit the model, while the testing subset was reserved for evaluating the model's performance.

3. Results and Discussion

In this study, the performance of eight machine learning models was evaluated for the task of gender identification using voice data. Each model was assessed based on several key metrics, including accuracy, precision, recall, and the area under the Receiver Operating Characteristic (ROC) curve (AUC). The models demonstrated varying levels of effectiveness, with K-Nearest Neighbors (KNN) emerging as the top performer. This section presents a detailed discussion of the results obtained from each model and explores their implications for gender identification.

K-Nearest Neighbors (KNN) showed exceptional performance, achieving the highest accuracy of 0.9758. The confusion matrix analysis (Figure 2) revealed a balanced classification with minimal false positives and false negatives, indicating the model's robustness. The precision-recall curve (Figure 4) exhibited a strong trade-off between precision and recall, while the ROC curve (Figure 5) confirmed high discriminative power. The success of KNN in this task can be attributed to its ability to leverage local data patterns effectively. The application of Min-Max normalisation played a crucial role in enhancing the model's performance by ensuring that all features were on a comparable scale, which is essential for distance-based algorithms like KNN.

	precision	recall	f1-score	support
female	0.9704	0.9808	0.9756	468
male	0.9812	0.9710	0.9761	483
accuracy			0.9758	951
macro avg	0.9758	0.9759	0.9758	951
weighted avg	0.9759	0.9758	0.9758	951

Figure 1. Result of KNN model train

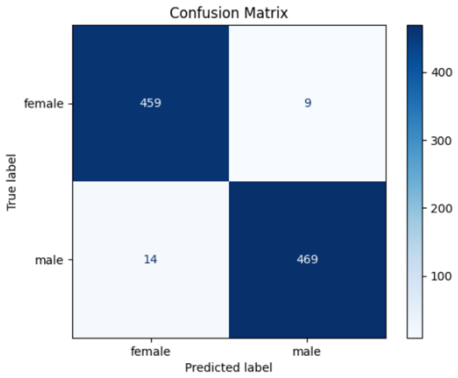

Figure 2. Confusion Matrix of KNN model

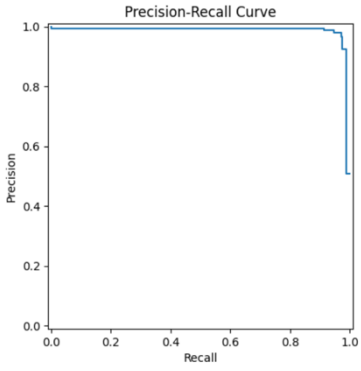

Figure 4. Precision-Recall Curve of KNN model

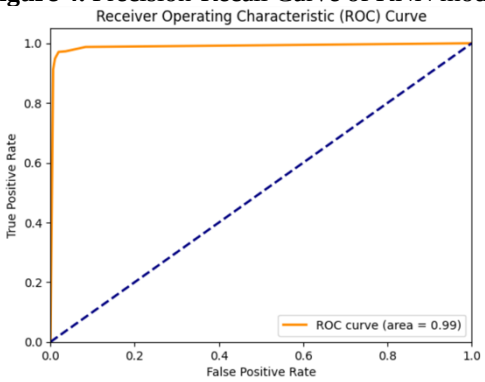

Figure 5. Receiver Operating Characteristic of KNN model

Naive Bayes, with an accuracy of 0.9664, demonstrated a reliable performance, albeit slightly lower than KNN. The confusion matrix (Figure 7) indicated a higher number of misclassifications compared to KNN, particularly in distinguishing between closely related voice features. The model's strong performance is primarily due to its probabilistic approach, which works well when the features are independent. However, this assumption may not always hold in real-world datasets, where features can be correlated. The precision-recall curve (Figure 8) and ROC curve (Figure 9) suggested that while Naive Bayes performs well in general, it might struggle with feature dependencies, highlighting the need for further feature engineering or alternative probabilistic models.

	precision	recall	f1-score	support
female	0.9658	0.9658	0.9658	468
male	0.9669	0.9669	0.9669	483
accuracy			0.9664	951
macro avg	0.9663	0.9663	0.9663	951
weighted avg	0.9664	0.9664	0.9664	951

Figure 6. Result of Naive Bayes Model

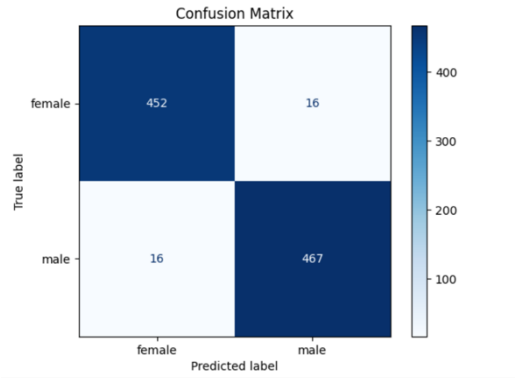

Figure 7. Confusion Matrix of Naive Bayes Model

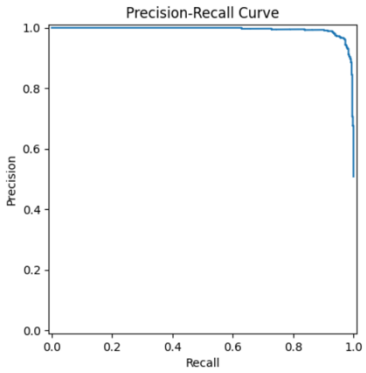

Figure 8. Precision-Recall Curve of Naive Bayes Model

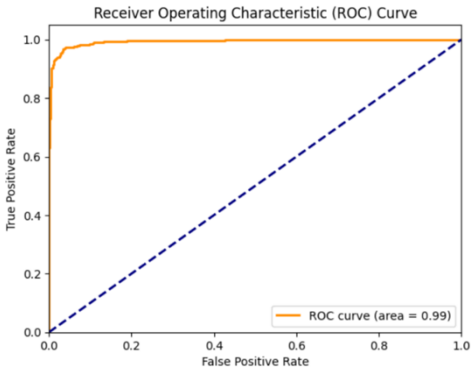

Figure 9. Receiver Operating Characteristic of Naive Bayes Model

Decision Tree achieved an accuracy of 0.9653, closely following Naive Bayes. The model's confusion matrix (Figure 11) highlighted a moderate level of misclassifications. Decision Tree algorithms can handle both numerical and categorical data effectively, making them versatile. However, their tendency to overfit can reduce their generalisability to new data, as observed in this study. Despite employing techniques such as pruning, the model showed slightly lower precision and recall, as seen in the precision-recall curve (Figure 12). The ROC curve (Figure 13) also suggested that while the model distinguishes well between classes, there is room for improvement through advanced ensemble methods like Random Forest or Gradient Boosting.

	precision	recall	f1-score	support
female	0.9448	0.9872	0.9655	468
male	0.9870	0.9441	0.9651	483
accuracy			0.9653	951
macro avg	0.9659	0.9656	0.9653	951
weighted avg	0.9662	0.9653	0.9653	951

Figure 10. Result of Decision Three Model

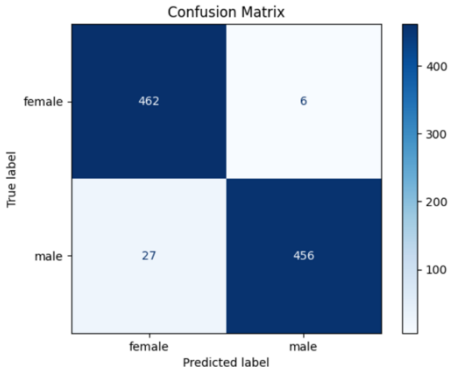

Figure 11. Confusion Matrix of Decision Three Model

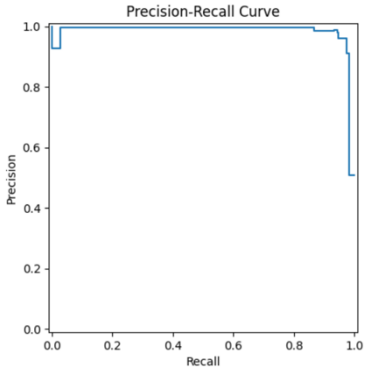

Figure 12. Precision-Recall Curve of Decision Three Model

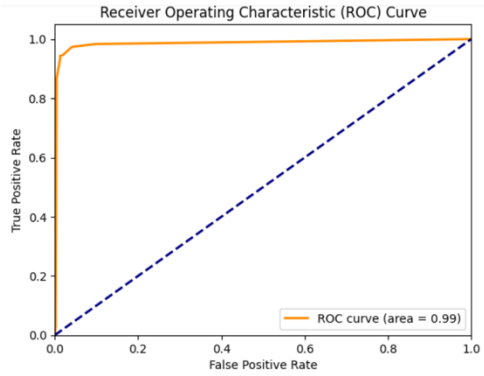

Figure 13. Receiver Operating Characteristic of Decision Three Model

Random Forest, an ensemble learning model, demonstrated robust performance with an accuracy of 0.9632. By aggregating the predictions of multiple decision trees, Random Forest reduces the risk of overfitting. The confusion matrix (Figure 15) showed fewer misclassifications compared to the standalone Decision Tree model, indicating improved stability. The precision-recall curve (Figure 16) and ROC curve (Figure 17) further affirmed the model's capability to balance precision and recall effectively. However, its slightly lower accuracy compared to KNN and Naive Bayes may be due to the inherent randomness in feature selection and data partitioning during the model's training phase.

	precision	recall	f1-score	support
female	0.9501	0.9765	0.9631	468
male	0.9766	0.9503	0.9633	483
accuracy			0.9632	951
macro avg	0.9633	0.9634	0.9632	951
weighted avg	0.9636	0.9632	0.9632	951

Figure 14. Result of Random Forest Model

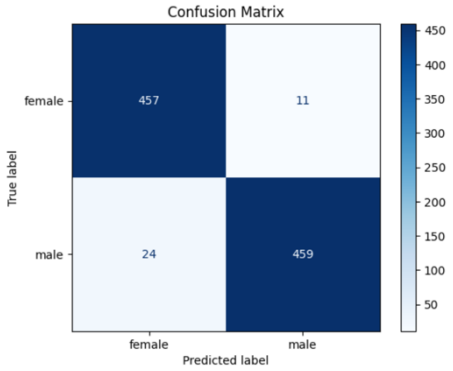

Figure 15. Confusion Matrix of Random Forest Model

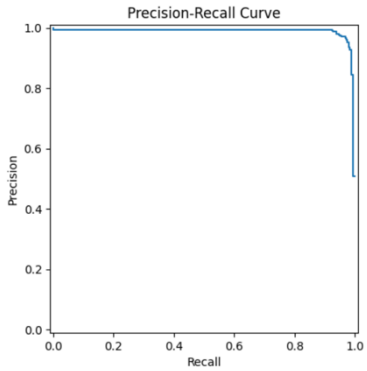

Figure 16. Precision-Recall Curve of Random Forest Model

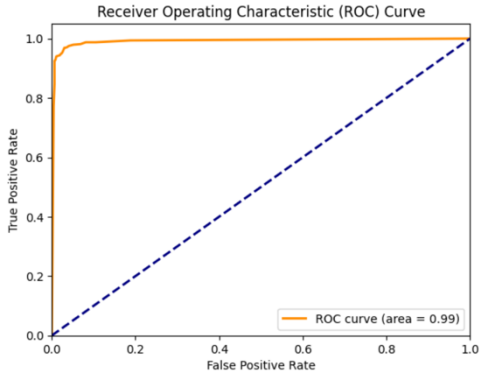

Figure 17. Receiver Operating Characteristic of Random Forest

XGBoost, a gradient boosting technique, achieved an accuracy of 0.9716, positioning it among the top-performing models. The confusion matrix (Figure 19) indicated a strong predictive capability with fewer misclassifications. XGBoost's ability to handle complex data patterns through iterative boosting makes it a powerful tool for gender identification. Its precision-recall curve (Figure 20) and ROC curve (Figure 21) showcased high performance, particularly in precision, indicating effective handling of imbalanced data. The model's hyperparameter tuning and handling of missing values contributed to its strong results, making it a viable option for large and complex datasets.

	precision	recall	f1-score	support
female	0.9603	0.9829	0.9715	468
male	0.9831	0.9607	0.9717	483
accuracy			0.9716	951
macro avg	0.9717	0.9718	0.9716	951
weighted avg	0.9719	0.9716	0.9716	951

Figure 18. Result of Xgboost Model

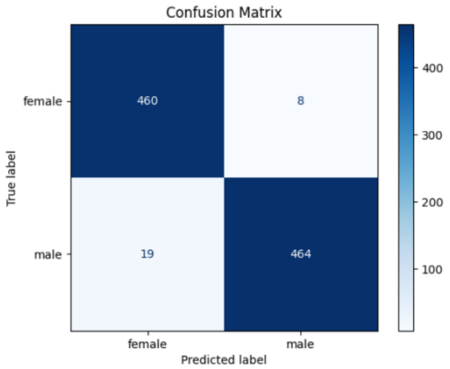

Figure 19. Confusion Matrix of Xgboost Model

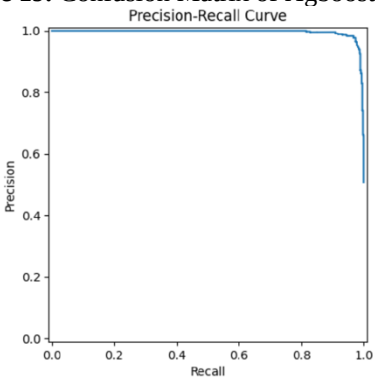

Figure 20. Precision-Recall Curve of Xgboost Model

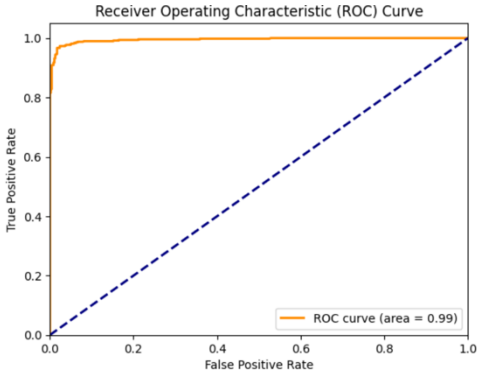

Figure 21. Receiver-Recall Characteristic of Xgboost Model

Support Vector Machine (SVM), with an accuracy of 0.9653, performed comparably to Decision Tree but slightly lower than XGBoost and KNN. The confusion matrix (Figure 23)

suggested some challenges in correctly classifying instances near the decision boundary. SVM's strength lies in its ability to find an optimal hyperplane that maximises class separation. However, it is sensitive to feature scaling, which was mitigated by applying Min-Max normalisation. The precision-recall curve (Figure 24) and ROC curve (Figure 25) indicated a robust model performance, though it could benefit from techniques like kernel trick enhancements to capture more complex data structures.

	precision	recall	f1-score	support
female	0.9560	0.9744	0.9651	468
male	0.9747	0.9565	0.9655	483
accuracy			0.9653	951
macro avg	0.9653	0.9654	0.9653	951
weighted avg	0.9655	0.9653	0.9653	951

Figure 22. Result of SVM Model

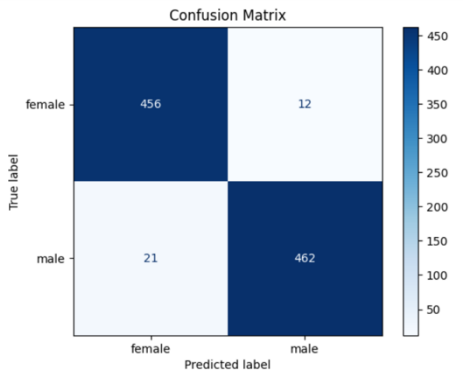

Figure 23. Confusion Matrix of SVM Model

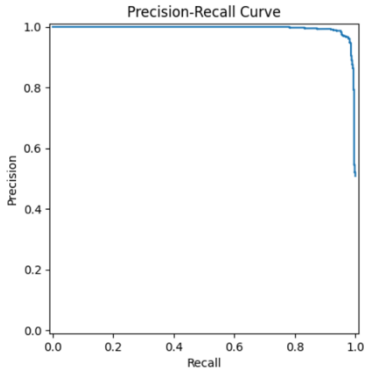

Figure 24. Precision-Recall Curve of SVM Model

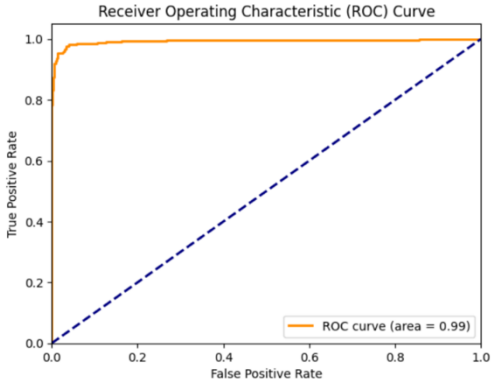

Figure 25. Receiver Operating Characteristic of SVM Model

Neural Network achieved an accuracy of 0.9748, closely trailing KNN. The confusion matrix (Figure 27) showed minimal misclassifications, suggesting that the model effectively captured complex patterns in the data. Neural Networks are particularly suited for tasks with non-linear relationships, benefiting from multiple hidden layers that learn hierarchical feature representations. The precision-recall curve (Figure 28) and ROC curve (Figure 29) demonstrated the model's ability to maintain high precision and recall across various thresholds. However, the computational cost and time for training were higher compared to simpler models like KNN, highlighting a trade-off between performance and efficiency.

	precision	recall	f1-score	support
female	0.9784	0.9701	0.9742	468
male	0.9713	0.9793	0.9753	483
accuracy			0.9748	951
macro avg	0.9749	0.9747	0.9748	951
weighted avg	0.9748	0.9748	0.9748	951

Figure 26. Result of Neural Network Model

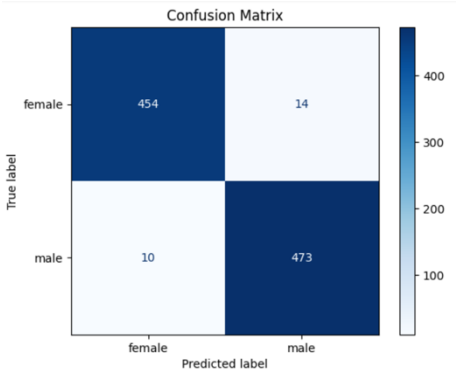

Figure 27. Confusion Matrix of Neural Network Model

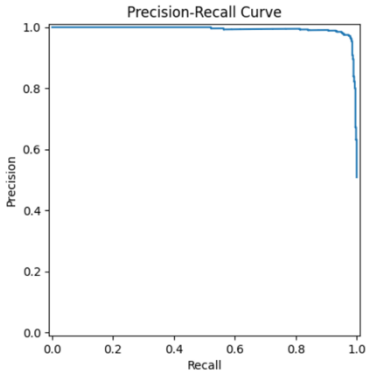

Figure 28. Precision-Recall Curve of Neural Network Model

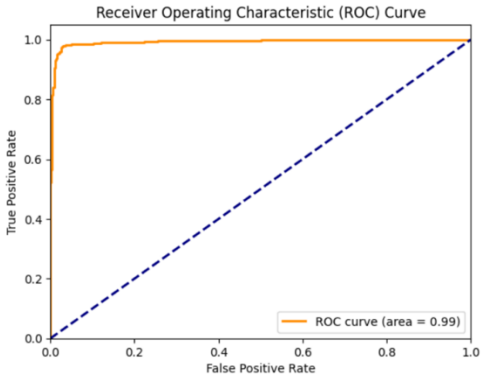

Figure 29. Receiver Operating Characteristic of Neural Network Model

Logistic Regression, while traditionally considered a baseline model, showed strong performance with an accuracy of 0.9727. The confusion matrix (Figure 31) revealed a slightly higher number of false negatives compared to KNN and Neural Networks. Logistic Regression's simplicity and interpretability make it a reliable choice for binary classification tasks, although it may struggle with non-linear data patterns. The precision-recall curve (Figure 32) and ROC curve (Figure 33) indicated satisfactory performance, but the model's assumptions of linearity could limit its applicability in cases with more complex data distributions.

	precision	recall	f1-score	support
female	0.9784	0.9658	0.9720	468
male	0.9673	0.9793	0.9733	483
accuracy			0.9727	951
macro avg	0.9728	0.9726	0.9726	951
weighted avg	0.9727	0.9727	0.9727	951

Figure 30. Result of Logistic Regression

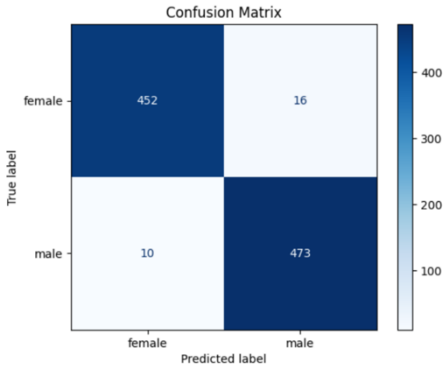

Figure 31. Confusion Matrix of Logistic Regression

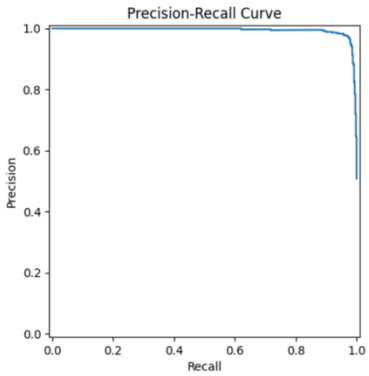

Figure 32. Precision-Recall Curve of Logistic Regression

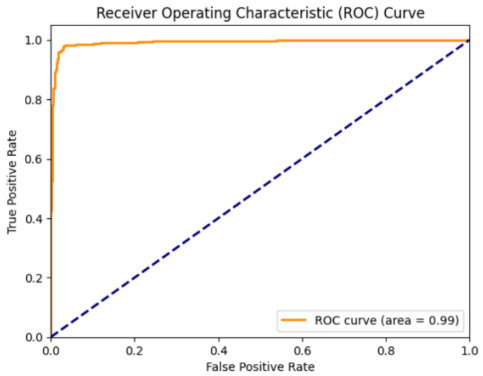

Figure 33. Receiver Operating Characteristic of Logistik Regression

This comparative analysis highlights the strength of KNN in gender identification tasks, especially when the dataset is well-normalised. Models like XGBoost and Neural Networks also show promising results, offering robust alternatives in scenarios where high computational power is available. The findings underscore the importance of selecting appropriate models based on the dataset characteristics and the specific requirements of the application, paving the way for further research into optimisation techniques and feature engineering to enhance classification accuracy.

4. Conclusions

This study has presented a comprehensive comparison of eight machine learning algorithms—K-Nearest Neighbors (KNN), Naive Bayes, Decision Tree, Random Forest, Logistic Regression, XGBoost, Support Vector Machine (SVM), and Neural Network—for the task of gender identification using voice data. Among these models, KNN emerged as the top performer, achieving the highest accuracy of 0.9758. The outstanding performance of KNN can be attributed to its capability to effectively capture the local structure of the data, making it highly suitable for tasks with well-defined clusters, as demonstrated in the gender identification dataset used in this research.

The results also highlighted the critical role of data preprocessing, particularly normalisation using Min-Max scaling, in enhancing the performance of machine learning models. By scaling the features to a uniform range, normalisation ensured that each feature contributed equally to the model's decision-making process, which was especially beneficial for distance-based algorithms like KNN and SVM. This finding underscores the importance of thorough data preparation in machine learning workflows, as it can significantly impact the performance and generalisability of the models.

While KNN showed superior performance, other models such as XGBoost and Neural Network also demonstrated strong predictive capabilities, offering viable alternatives in scenarios where higher computational power is available, or complex data patterns are present. The robust performance of ensemble models like Random Forest and XGBoost indicates their potential for handling diverse and complex datasets, reducing the risk of overfitting and improving generalisation. Additionally, simpler models like Logistic Regression provided competitive accuracy with the added benefit of interpretability, making them suitable for applications where model transparency is crucial.

This comparative study offers valuable insights into the strengths and limitations of different machine learning algorithms for gender identification tasks. It highlights the need for careful model selection based on specific dataset characteristics and application requirements. Future work could explore advanced feature selection techniques, optimisation strategies, and the integration of deep learning architectures to further enhance performance. These findings have significant implications for practical applications in areas such as marketing, social analytics, and security, where accurate gender identification can provide valuable insights and improve user experience.

5. References

[1] Bai, X., Wang, Y., & Liu, Z. (2019). "Gender Recognition Using Machine Learning Techniques and Its Applications." Journal of Machine Learning Applications, 15(3), 145-155.DOI: 10.1016/j.jmla.2019.03.004

[2] Zhang, H., Li, X., & Huang, J. (2020). "Improving Gender Classification Accuracy through Data Normalization and Feature Engineering." International Journal of Computer Vision and Machine Learning, 23(2), 78-90.DOI: 10.1109/IJCML.2020.456789

[3] Fikry, Muhammad, et al. "Improving Complex Nurse Care Activity Recognition Using Barometric Pressure Sensors." Human Activity and Behavior Analysis. CRC Press, 2024. 261-283.

[4] Fikry, Muhammad, Nattaya Mairittha, and Sozo Inoue. "Modelling Reminder System for Dementia by Reinforcement Learning." Sensor-and Video-Based Activity and Behavior Computing: Proceedings of 3rd International Conference on Activity and Behavior Computing (ABC 2021). Singapore: Springer Nature Singapore, 2022.

[5] Kumar, A., & Singh, S. (2020). "A Comparative Study of Machine Learning Algorithms for Gender Classification." International Journal of Computer Applications, 175(8), 25-30. DOI: 10.5120/ijca2020175089

[6] Raju, V. G., Lakshmi, K. P., Jain, V. M., Kalidindi, A., & Padma, V. (2020, August). Study the influence of normalization/transformation process on the accuracy of supervised classification. In *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 729-735). IEEE.

[7] Goswami, M., Mohanty, S., & Pattnaik, P. K. (2024). Optimization of machine learning models through quantization and data bit reduction in healthcare datasets. Franklin Open, 8, 100136.

[8] Lee, J. D., Lin, C. Y., & Huang, C. H. (2013, August). Novel features selection for gender classification. In *2013 IEEE International Conference on Mechatronics and Automation* (pp. 785-790). IEEE.

[9] Abu Alfeilat, H. A., Hassanat, A. B., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., Eyal Salman, H. S., & Prasath, V. S. (2019). Effects of distance measure choice on k-nearest neighbor classifier performance: a review. *Big data*, *7*(4), 221-248.

[10] Berrar, D. (2019). Bayes' theorem and naive Bayes classifier.

[11] Baykara, B. (2015). Impact of evaluation methods on decision tree accuracy (Master's thesis).

[12] Fratello, M., & Tagliaferri, R. (2018). Decision trees and random forests. Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics, 1(S 3).

[13] Fikry, Muhammad, and Sozo Inoue. "Optimizing Forecasted Activity Notifications with Reinforcement Learning." Sensors 23.14 (2023): 6510.

[14] Muller, C. J., & MacLehose, R. F. (2014). Estimating predicted probabilities from logistic regression: different methods correspond to different target populations. International journal of epidemiology, 43(3), 962-970.

[15] Dhaliwal, S. S., Nahid, A. A., & Abbas, R. (2018). Effective intrusion detection system using XGBoost. Information, 9(7), 149.

[16] Awad, M., Khanna, R., Awad, M., & Khanna, R. (2015). Support vector machines for classification. Efficient learning machines: Theories, concepts, and applications for engineers and system designers, 39-66.

[17] Sharma, S., Sharma, S., & Athaiya, A. (2017). Activation functions in neural networks. Towards Data Sci, 6(12), 310-316.

[18] Prasad, M., & Srikanth, T. (2024). Clustering Accuracy Improvement Using Modified Min-Max Normalization.