Bulletin of Informatics and Data Science
Vol. 4 No. 1, May 2025, Page 34−41
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.115
https://ejurnal.pdsi.or.id/index.php/bids/index

Implementation of XGBoost Ensemble and Support Vector Machine For Gender Classification of Skull Bones

Astrid Ramadhani, Iis Afrianty*, Elvia Budianita, Siska Kurnia Gusti

Faculty of Science and Technology, Informatics Engineering, Sultan Syarif Kasim Riau State Islamic University, Pekanbaru, Indonesia
Email: [1]12150121537@students.uin-suska.ac.id, [2,*]iis.afrianty@uin-suska.ac.id, [3]elvia.budianita@uin-suska.ac.id, [4]siskakurniagusti@uin-suska.ac.id
Correspondence Author Email: iis.afrianty@uin-suska.ac.id

Abstract

Sex identification based on skull bones is an important step in forensic anthropology, especially in cases where unidentified human skeletons are found. Conventional methods such as DNA analysis are often used, but have limitations, especially when the bones are damaged, charred or decayed, making the analysis process difficult. This research applies XGBoost ensemble and Support Vector Machine for sex classification on skull bones. The purpose of this research is to handle complex data with many features and unbalanced data using the XGBoost ensemble method and Support Vector Machine (SVM). The data used consisted of 2,524 samples with 82 measurement features. Model performance was evaluated using accuracy, precision, recall, and F1 score metrics. The results showed that the combination of XGBoost and SVM methods, especially with the RBF kernel, was able to achieve accuracy of up to 91.52%. This finding proves that machine learning-based approaches can be an effective and reliable solution in supporting the forensic identification process.

Keywords: Forensic Anthropology; Support Vector Machine; Skull; Voting Classifier; XGBoost

1. INTRODUCTION

Forensic anthropology is the application of anthropological methods and theories in skeletal recovery and analysis to understand human variation, skeletal biology, and bone biomechanics [1]. By analyzing skeletal remains, the field can reconstruct the biological profile of a deceased person, including sex, age at death, and stature [1], [2], [3]. Sex identification is the first step in skeletal remains recognition and forms the basis for further analysis [1], [4], [5]. The skull is one of the most useful skeletal parts in sex classification, with an accuracy rate of up to 90% after the pelvic bones [3], [6], [7]. In addition, forensic anthropology methods based on metric measurements are known to have a high accuracy rate and can be performed quickly if bone conditions permit and the observer has sufficient experience [1], [3], [6].

In an attempt to determine the sex of skeletal remains, various methods can be used, such as DNA analysis in the laboratory and radiological examination [3], [5], [8]. However, these methods have limitations under certain conditions, such as if the bones have been burnt or damaged making DNA extraction difficult [1], [8], [9]. In addition, laboratory methods are generally costly and time-consuming as they require special preparation and in-depth examination of the skeleton [10]. To overcome these obstacles, machine learning-based approaches are being used in sex classification from skeletal remains [2], [8].

One of the frequently used machine learning techniques is ensemble learning, which improves prediction accuracy by combining several individual models [10], [11]. Ensemble learning is a technique in machine learning that combines prediction results from several algorithms to improve classification performance [12]. This technique utilizes the power of various models to get more accurate results than a single method [10], [11], [13]. In previous research, boosting-based ensemble methods such as XGBoost, AdaBoost, and Gradient Boosting have shown improved accuracy in various classification tasks [12]. One approach that can be applied in forensics is the combination of XGBoost and SVM methods to increase the effectiveness in identifying the gender of skull bones [10], [14].

The Support Vector Machine (SVM) method itself is also used in gender classification from skull bone data and has been shown to have high performance [6]. SVM models with Radial Basis Function (RBF) kernels yielded accuracies above 90% in various forensic-related studies [6]. However, some studies have shown that other algorithms, such as XGBoost, are capable of providing superior results in certain classification tasks [10]. XGBoost is an extension of the gradient boosting technique designed to improve the performance of classification and regression models by reducing complexity and preventing overfitting [15]. In a study related to breast cancer detection, XGBoost with feature selection was able to achieve an accuracy of 91.4%, higher than SVM which only reached 89.8% [13].

In addition to improving classification accuracy, the Adaptive Synthetic Sampling (ADASYN) technique can be used to handle data imbalance in the sex classification process [16], [17]. ADASYN works by adaptively generating synthetic samples for minority classes based on the distribution of the data, thereby increasing the sensitivity of the model in recognizing patterns from those classes [18]. In a study related to lung cancer, the combination of ADASYN and SVM successfully increased the prediction accuracy to 98.95% compared to the model without ADASYN which only reached 59% [18]. This technique is very useful in a forensic context, especially when the available data is not balanced between male and female gender [16], [17], [19]. By utilizing ADASYN, machine learning models can be more effective in classifying gender based on skull characteristics. Therefore, the application of ADASYN in forensic classification can help improve the accuracy of gender identification [16], [19].

Bulletin of Informatics and Data Science
Vol. 4 No. 1, May 2025, Page 34−41
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.115
https://ejurnal.pdsi.or.id/index.php/bids/index

Based on the description above, this research will be conducted by applying the ensemble learning method using a combination of XGBoost and SVM for sex classification based on skull bones. This research aims to evaluate the performance of the method in supporting the forensic identification process and improving the accuracy of sex classification. By integrating the ADASYN technique for data balancing, it is expected that the developed model can be more accurate and reliable in skull bone classification. The results of this research also have the potential to make a significant contribution to the field of forensics, particularly in victim identification based on skeletal remains. In addition, this approach can be applied to various other fields that require machine learning-based classification under conditions of unbalanced data.

2. RESEARCH METHODOLOGY

Research methodology is a series of interrelated and continuous processes or stages. These stages are outlined in a research method that is clearly described, structured, and systematic. The following explanation outlines the stages of this final project research, which guides the author in completing the research. The scheme of the research methodology used can be seen in Figure 1 as a visual guide.

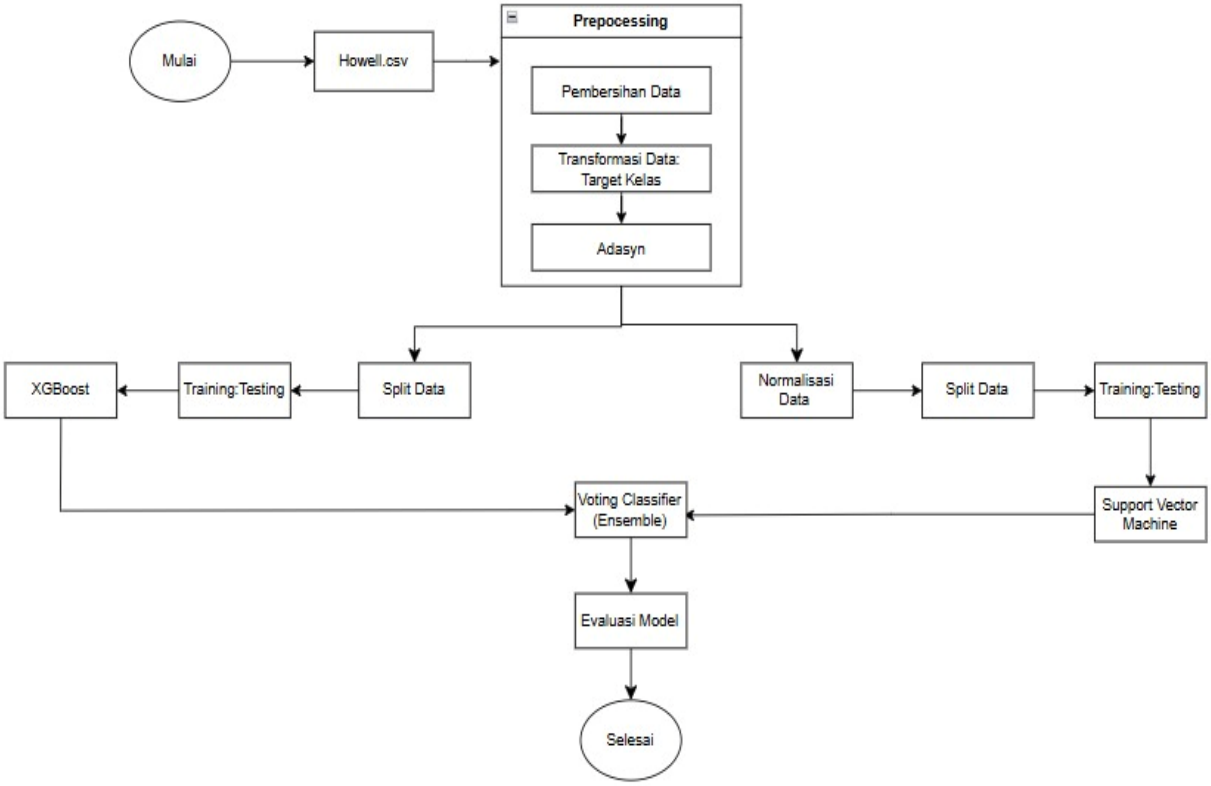

Figure 1. Schematic of the research methodology

2.1 Dataset Collection

This research process begins with the first step, which is to utilize a skull bone dataset that consists of information for skull bone type classification. The data used is the Howell.csv dataset, containing craniometric measurements collected by Dr. William W. Howells between 1965 and 1980. The dataset includes 2,524 human skull data, consisting of 1,368 male skulls and 1,156 female skulls, with parameter features such as Biauricular Breadth (AUB), Parietal Subtense (PAS), and Nasal Height (NLH), which are described in Table 1 and Table 2. The secondary data is taken from the website https://web.utk.edu/~auerbach/HOWL.htm, focusing on 82 features as sex-determining metrics, which are classified into two classes: male and female.

Table 1. Five Features of Skull Bone Parameters

Code	Features of Skull Bone Parameters
AUB	Biauricular breadth
PAS	Parietal Subtense
FRS	Frontal Subtense
...	...
TBA	Thiobarbituric acid

Bulletin of Informatics and Data Science
Vol. 4 No. 1, May 2025, Page 34−41
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.115
https://ejurnal.pdsi.or.id/index.php/bids/index

Table 2. Skull Bone Measurements

CLASS	GOL	ZYB	AUB	RFA	OCA	...	TBA
M	189	133	119	0	117	...	0
M	182	137	125	0	119	...	0
F	156	113	102	63	133	...	151
...	...	...	...	...	...	...	...
F	160	117	112	60	131	...	156

2.2 Data Preprocessing

The next step is data selection, where irrelevant features such as ID, PopNum, and Population are removed to increase the relevance of the data in the classification process. After data selection, data *preprocessing* is performed to clean the data from noise or missing values and ensure the data is ready for further analysis stages. This stage is followed by data transformation, which converts non-numeric data such as text into numeric data. In the context of this study, gender labels were transformed, with males coded as 0 and females as 1.

This research process involves several important stages designed to produce an accurate model for gender classification based on skull bone craniometric features. After the preprocessing and data transformation stages are completed, the next step is to apply the ADASYN *(Adaptive Synthetic Sampling)* technique. This technique is used to handle class imbalance in the data. Class imbalance often occurs when one category in a dataset is much more numerous than another, which can result in the model tending to predict the majority category. ADASYN works by adaptively synthesizing new data samples from minority classes based on the distribution of the data, thus helping to improve model performance on underrepresented classes.

2.3 Data Normalization

Data normalization is done specifically for data that will be processed by the *Support Vector Machine* (SVM) algorithm. This normalization aims to ensure that all features in the dataset have a uniform scale, considering that the SVM algorithm is sensitive to differences in scale between features. The min-max normalization formula is shown in Equation 1 [20].

$$'\mathrm{X_i} = \frac{\mathrm{X_i} - \min_{(x)}}{\max_{(x)} - \min_{(x)}} (\max_{(baru)} - \min_{(baru)}) + \min_{(baru)} \quad (1)$$

a. X_i = states the original value
b. $'X_i$ = value after normalization
c. $max_{(x)} - min_{(x)}$ = minimum and maximum values of the change in X
d. $max_{(baru)} - min_{(baru)}$ = new desired range

2.4 Split Data

After the ADASYN and Data Normalization techniques are applied, the data is divided into two main parts: *training data* and *testing data*. *Training data* is used to train the model to understand the patterns in the data, while *testing data* is used to evaluate the model's performance and ensure that the model can give good predictions on new data that has never been seen before. This division of data is important to prevent overfitting, where the model only works well on training data but fails on new data. In this study, the data splitting was carried out using three different ratios: 70:30, 80:20, and 90:10, to evaluate the impact of training data size on model performance.

2.5 Training Single Models

Support Vector Machine (SVM) can be used as an analytical tool to classify or identify human bone traits, such as gender, age, or population of origin, based on morphological features or bone metrics. SVM works by finding the optimal hyperplane that separates bone data into specific categories, such as male and female, or specific age groups. In cases where the difference between categories is not linear, SVM uses a kernel trick to map the data to a higher dimensional space, where linear separation becomes possible. For example, in determining gender based on pelvic bone size, SVM can maximize the margin between two classes (male and female) by considering the *support vectors*, which are the bone data points closest to the *hyperplane*. Thus, SVM helps anthropology in making accurate and objective predictions, especially when facing complex or highly variable bone data. The following kernel functions are used in Table 3 [21].

Table 3. Kernel Function

Kernel Name	Kernel Function	
Linear	$K(x_i, x_j) = x_j., x_j$	(2)
RBF	$K(x_i, x_j) = (x_j., x_j + c)^d$	(3)
Polynomial	$K(x_i, x_j) = exp(-y\|\|x_j - x_j\|\|^d$	(4)
Sigmoid	$K(xi, xj) = tanh\ (yx_i^T x + r)$	(5)

Bulletin of Informatics and Data Science
Vol. 4 No. 1, May 2025, Page 34−41
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.115
https://ejurnal.pdsi.or.id/index.php/bids/index

XGBoost (*Extreme Gradient Boosting*) is an effective machine learning algorithm for classification tasks, including in the field of bone anthropology, specifically for classifying gender based on bone characteristics. The algorithm works by sequentially building a series of decision trees, where each tree attempts to correct the prediction error of the previous tree through boosting techniques. In the context of sex classification, bone features such as femur length, femur head diameter, or pelvis size are used as inputs to the model. XGBoost optimizes hyperparameters such as tree depth (*max_depth*), *learning* rate, and number of trees (*n_estimators*) to minimize the prediction error. Once the model is trained, its performance is evaluated using metrics such as accuracy, precision, recall, and F1-score. The trained model can then be used to predict gender based on new bone data. In addition, XGBoost provides an importance feature that allows anthropologists to understand which bone characteristics are most influential in distinguishing sex. With its high accuracy and ability to handle complex data, XGBoost is a very useful tool in bone anthropology analysis.

2.6 Voting Classifier

In this research, XGBoost and SVM algorithms are used to build predictive models. The prediction results from the two models are then combined using the Ensemble Voting Classifier method, which aims to improve prediction accuracy by combining the advantages of each algorithm. This ensemble technique is often used to produce more stable and accurate predictions than using a single model. Specifically, in this study, four different Voting Classifiers were built, each combining XGBoost with one type of SVM kernel (Linear, RBF, Polynomial, and Sigmoid) separately, by applying soft voting strategies.

2.7 Evaluation Models

The final stage in this research is model evaluation using testing data. Evaluation is done to measure the overall performance of the model based on confusion metrics, such as accuracy, precision, and recall. The evaluation results become the basis for assessing the extent to which the model is successful in gender classification based on craniometric data.

3. RESULT AND DISCUSSION

The experiment was conducted using 2524 skull bones with 82 variables that have been implemented through Python programming by utilizing tools from Google Colab. These 82 variables are the result of the data cleaning process carried out in the preprocessing stage. Experiments were conducted using four types of kernels in Support Vector Machine (SVM), namely linear, radial basis function (rbf), polynomial, and sigmoid. Among the four, the SVM kernel model that is able to evaluate the model well can be seen from the accuracy results measured using Confusion Matrix. The accuracy results obtained from each SVM kernel can be seen in the table provided.

The linear SVM kernel is a type of kernel function that is effective when the data is linearly separable. This kernel is suitable when the data is naturally linearly separable. Linear kernels are also very suitable for data with high feature dimensions, as mapping to higher dimensional spaces does not provide a significant performance improvement. In the code implementation, the parameters used are C=10, gamma=1, and probability=True. The value of parameter C controls the balance between the smoothness of the decision boundary and the classification accuracy of the training points. In addition, a sigmoid kernel was also tested, which is suitable for cases where the data has non-linear characteristics and requires transformation to a different feature space. The sigmoid kernel is often used in binary classification problems and can give good results if its parameters are set appropriately.

In addition to SVM, training is also carried out using the XGBoost algorithm with the best parameters that have been determined. The best parameters used in XGBoost training are as follows: objective set to 'binary:logistic' for binary classification, eval_metric using 'logloss', max_depth set to 3, learning_rate to 0.1 and 0.01, and n_estimators to 300. With these parameters, the XGBoost model was trained and tested for accuracy. The accuracy results of the XGBoost model are also reported for further evaluation, showing competitive performance compared to the SVM model. This combination of using various SVM kernels and XGBoost allows a thorough evaluation of the model's performance in handling complex data. The linear kernel accuracy comparison can be seen in Table 4.

Table 4. Accuracy of Ensemble Voting Classifier Model (SVM with Linear kernel and XGBoost) Based on Data Ratio and Learning Rate XGBoost

Ratio	Learning Rate XGBoost	Voting Classifier Accuracy
70:30	0.1	88.31%
	0.01	87.03%
80:20	0.1	89.40%
	0.01	88.05%
90:10	0.1	88.04%
	0.01	88.07%

The RBF (Radial Basis Function) kernel is a widely used function due to its ability to handle data that cannot be linearly separated. This kernel has two main parameters, namely cost (C) and gamma (γ). The gamma parameter regulates

Bulletin of Informatics and Data Science
Vol. 4 No. 1, May 2025, Page 34−41
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.115
https://ejurnal.pdsi.or.id/index.php/bids/index

how much influence a data point has in the training process, where a low gamma value indicates a wider influence, while a high gamma value indicates a more local influence. If gamma is too small, the model becomes too simple and unable to capture the complexity of the data. Conversely, if gamma is too large, the model may become too specific and prone to overfitting. The C parameter serves to control regularization, helping to prevent overfitting by balancing the margin and classification error.

In the code implementation, the gamma parameter is set to 1, while the C parameter is set to 10. The value of C=10 indicates that the model has a higher tolerance to misclassification during training, making it more flexible. In addition, the *probability=True* parameter is used to allow the model to generate probability estimates during the prediction process. The accuracy results of the SVM model with these parameters can be seen in the table provided.

In addition to SVM, training was also conducted using the XGBoost algorithm with the best predefined parameters. The parameters used include *objective='binary:logistic'* for binary classification, eval_metric='logloss' as the evaluation metric, *max_depth=3* to limit the maximum depth of the tree, *learning_rate=0.*1 and 0.01 to control the learning speed, and *n_estimators=300* to determine the number of trees used. With these parameters, the XGBoost model was trained and tested to produce accuracy. The accuracy results of the XGBoost model are also reported, showing competitive performance when compared to the SVM model. The combined use of these two models allows for a more comprehensive evaluation of the model's performance in handling complex data. The rbf kernel accuracy comparison can be seen in Table 5.

Table 5. Accuracy of Ensemble Voting Classifier Model (SVM with RBF kernel and XGBoost) Based on Data Ratio and Learning Rate XGBoost

Ratio	Learning Rate XGBoost	Voting Classifier Accuracy
70:30	0.1	88.83%
	0.01	88.96%
80:20	0.1	91.13%
	0.01	91.52%
90:10	0.1	90%
	0.01	89.61%

Polynomial kernel is one of the kernel functions used in Support Vector Machine (SVM) to handle non-linearly separable data. It works by mapping the data to a higher feature space through a polynomial of a certain degree, thus allowing the model to capture non-linear relationships between features. The main parameters in a polynomial kernel are *degree*, which determines the degree of the polynomial, as well as gamma which controls the influence of each data point. Polynomial kernels are particularly useful when the data has a complex structure and requires a non-linear transformation to achieve optimal separation. However, choosing too high a polynomial degree may lead to overfitting, while too low a degree may not be enough to capture the complexity of the data.

In the code implementation, the degree parameter is set to 3, indicating that the model will use a third-degree polynomial to transform the data. The gamma parameter is set to 1, while the C parameter is set to 10, providing greater tolerance to misclassification during training, making the model more flexible. In addition, the *probability=True* parameter is used to allow the model to generate probability estimates during the prediction process. The accuracy results of the SVM model with these parameters can be seen in the table provided.

On the other hand, training is also done using the XGBoost algorithm with the best predefined parameters. Some of the parameters used include objective=*'binary:logistic'* for binary classification, eval_metric='logloss' as the evaluation metric, *max_depth=3* to limit the maximum depth of the tree, learning_rate=0.1 and 0.01 to control the learning speed, and *n_estimators=300* to determine the number of trees used. With these parameters, the XGBoost model was trained and tested to produce accuracy. The accuracy results of the XGBoost model are also reported, showing competitive performance when compared to the SVM model. The combined use of these two models allows for a more comprehensive evaluation of the model's performance in handling complex data. The polynomial kernel accuracy comparison can be seen in Table 6.

Table 6. Accuracy of Ensemble Voting Classifier Model (SVM with Polynomial kernel and XGBoost) Based on Data Ratio and Learning Rate XGBoost

Ratio	Learning Rate XGBoost	Voting Classifier Accuracy
70:30	0.1	89.21%
	0.01	87.67%
80:20	0.1	90.36%
	0.01	88.63%
90:10	0.1	88.07%
	0.01	86.54%

The sigmoid kernel is one of the kernel methods in Support Vector Machine (SVM) designed to process data that is not linearly separable. It uses a sigmoid activation approach, similar to that of artificial neural networks. The sigmoid kernel is particularly effective for data that has non-linear characteristics, which require mapping to different feature

Bulletin of Informatics and Data Science
Vol. 4 No. 1, May 2025, Page 34−41
ISSN 2580-8389 (Media Online)
DOI 10.61944/bids.v4i1.115
https://ejurnal.pdsi.or.id/index.php/bids/index

spaces. However, the success of the sigmoid kernel is highly dependent on setting parameters such as gamma. Without careful parameter adjustment, the model can suffer from overfitting or underfitting. For implementation in code, the parameters are the same as the linear and rbf kernels. The sigmoid kernel accuracy comparison can be seen in Table 7.

Table 7. Accuracy of Ensemble Voting Classifier Model (SVM with Sigmoid kernel and XGBoost) Based on Data Ratio and Learning Rate XGBoost

Ratio	Learning Rate XGBoost	Voting Classifier Accuracy
70:30	0.1	87.68%
	0.01	82.93%
80:20	0.1	87.67%
	0.01	84.97%
90:10	0.1	85.77%
	0.01	82.31%

Based on the research results, the kernel with the highest performance is the RBF kernel. Visualization of the accuracy results of the Ensemble Voting Classifier model (using SVM with RBF and XGBoost kernels) based on data ratio and XGBoost learning rate is shown in Figure 2. Meanwhile, the evaluation of model performance through confusion matrix can be seen in Figure 3.

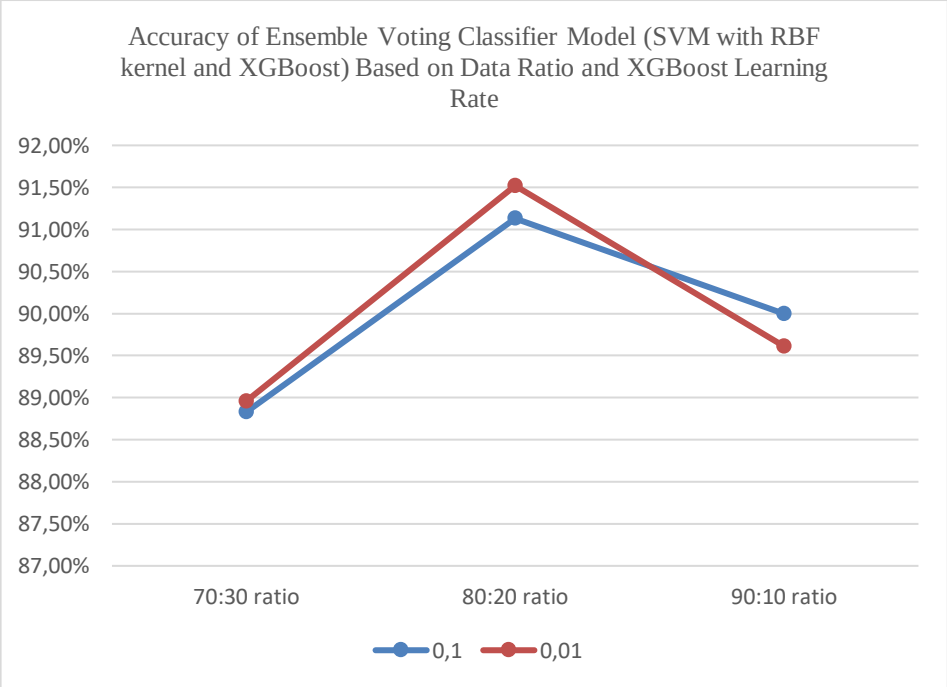

Figure 2. Accuracy of Ensemble Voting Classifier Model

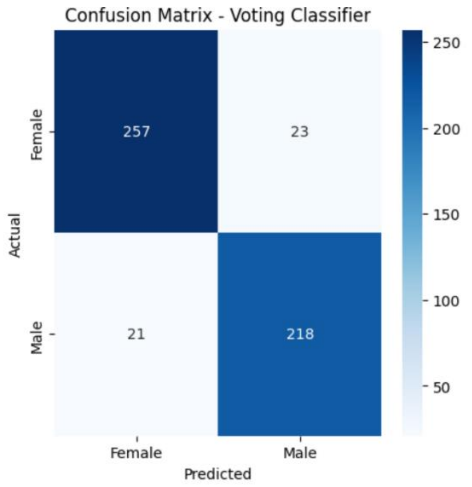

Figure 3. Confusion Matrix

4. CONCLUSION

Based on the results, the application of various kernels to the SVM algorithm as well as the use of XGBoost showed good classification performance on skull bone data. The RBF kernel provides the highest accuracy among all SVM kernels, proving its effectiveness in handling non-linear data. The polynomial kernel also provided competitive results, especially when compared to the linear and sigmoid kernels. This research aims to handle complex data with many features as well as imbalanced data by using the XGBoost ensemble method and Support Vector Machine (SVM), where its application is proven to improve classification performance compared to single methods. The highest classification accuracy achieved was 91.52%, obtained at a data ratio of 80:20, indicating a strong generalization ability on the test data. This improvement was obtained through the combination of appropriate parameters, such as data ratio, learning rate, and kernel configuration, which proved to contribute significantly to the accuracy improvement. This research confirms the importance of selecting the right model and parameters to achieve optimal classification results. For future research, it is recommended to apply feature selection techniques such as Gain Ratio or other approaches to improve the efficiency and accuracy of classification models. In addition, the research can also be continued by retraining using new data. This is important considering that the current research has only reached the modeling stage. In particular, it is recommended to retrain the meta-model in the stacking method, so that the model has better generalization ability to new data.

REFERENCES

[1] S. Aditya, I. Afrianty, S. Sanjaya, R. Abdillah, L. Handayani, and F. Insani, "Perbandingan Performansi Dengan Metode Correlation Based Feature Selection Pada LVQ 2," *Jurnal Instek*, vol. 8, no. 1, 2023.

[2] J. Bewes, A. Low, A. Morphett, F. D. Pate, and M. Henneberg, "Artificial intelligence for sex determination of skeletal remains: Application of a deep learning artificial neural network to human skulls," *J Forensic Leg Med*, vol. 62, pp. 40–43, Feb. 2019, doi: 10.1016/j.jflm.2019.01.004.

[3] Y. Harni, I. Afrianty, S. Sanjaya, R. Abdillah, F. Yanto, and F. Syafria, "Performance Analysis of LVQ 1 Using Feature Selection Gain Ratio for Sex Classification in Forensic Anthropology," *Building of Informatics, Technology and Science (BITS)*, vol. 5, no. 1, Jun. 2023, doi: 10.47065/bits.v5i1.3625.

[4] Darmila, I. Afrianty, S. Sanjaya, R. Abdillah, I. Iskandar, And F. Syafria, "Evaluasi Perbandingan Performansi Lvq 1, Lvq 2, Dan Lvq 3 Dalam Klasifikasi Jenis Kelamin Menggunakan Tulang Tengkorak," *Jurnal Instek*, vol. 7, no. 2, 2022.

[5] I. Afrianty, D. Nasien, and H. Haron, "Performance Analysis of Support Vector Machine in Sex Classification of The Sacrum Bone in Forensic Anthropology," *Jurnal Teknik Informatika*, vol. 15, no. 1, pp. 63–72, Jun. 2022, doi: 10.15408/jti.v15i1.25254.

[6] S. Sri Rahayu *et al.*, "Klasifikasi Tulang Tengkorak Berdasarkan Jenis Kelamin Dalam Antropologi Forensik Menggunakan Metode Support Vector Machine," *Jurnal Inovtek Polbeng -Seri Informatika*, vol. 9, no. 1, 2024.

[7] D. Nasien, M. Hasmil Adiya, I. Afrianty, N. A. Ali, A. A. Samah, and Y. Rahayu, "Determination of Sex and Race in Forensic Anthropology: A Comparison of Artificial Neural Network and Support Vector Machine," in *Proceedings - 2021 4th International Conference on Computer and Informatics Engineering: IT-Based Digital Industrial Innovation for the Welfare of Society, IC2IE 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 51–55. doi: 10.1109/IC2IE53219.2021.9649182.

[8] P. Mesejo, R. Martos, Ó. Ibáñez, J. Novo, and M. Ortega, "A Survey on artificial intelligence techniques for biomedical image analysis in skeleton-based forensic human identification," Jul. 01, 2020, *MDPI AG*. doi: 10.3390/app10144703.

[9] E. Nikita and P. Nikitas, "On the use of machine learning algorithms in forensic anthropology," *Leg Med*, vol. 47, Nov. 2020, doi: 10.1016/j.legalmed.2020.101771.

[10] D.- Andriansyah and Eka Wulansari Fridayanthie, "Optimization of Support Vector Machine and XGBoost Methods Using Feature Selection to Improve Classification Performance," *Journal Of Informatics And Telecommunication Engineering*, vol. 6, no. 2, pp. 484–493, Jan. 2023, doi: 10.31289/jite.v6i2.8373.

[11] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized XGBoost based diagnostic system for effective prediction of heart disease," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 7, pp. 4514–4523, Jul. 2022, doi: 10.1016/j.jksuci.2020.10.013.

[12] M. Rama Hadi Suryanto and D. Wahyu Utomo, "Pembelajaran Ensemble Untuk Klasifikasi Ulasan Pelanggan E-commerce Menggunakan Teknik Boosting," Infotekmesin, vol. 15, no. 02, 2024, doi: 10.35970/infotekmesin.v15i2.2314.

[13] M. Ravly Andryan *et al.*, "Komparasi Kinerja Algoritma Xgboost Dan Algoritma Support Vector Machine (Svm) Untuk Diagnosa Penyakit Kanker Payudara," *Jurnal Informatika dan Komputer)*, vol. 6, no. 1, pp. 1–5, 2022.

[14] R. Setiawan Aji Nugroho, "Comparison Of Support Vector Machine(Svm), Xgboost And Random Forest For Sentiment Analysis Of Bumble App User Comments," Jurnal Informatika Proxies, vol. 6, no. 1, 2022.

[15] I. Hanif, "Implementing Extreme Gradient Boosting (XGBoost) Classifier to Improve Customer Churn Prediction," European Alliance for Innovation n.o., Jan. 2020. doi: 10.4108/eai.2-8-2019.2290338.

[16] M. Tayebi and S. El Kafhali, "Generative Modeling for Imbalanced Credit Card Fraud Transaction Detection, "*Journal of Cybersecurity and Privacy*, vol. 5, no. 1, Mar. 2025, doi: 10.3390/jcp5010009.

[17] N. P. Kha, "Optimasi Metode CART Menggunakan Metode Bagging Pada Studi Kasus Data Imbalance Berbasis Metode Adasyn," Jurnal Ilmiah Matematika, vol. 10, no. 1, pp. 34–42, 2023, doi: 10.26555/konvergensi.30874.

[18] M. Tiara, T. B. Sirait, N. S. Fathonah, and M. N. Fauzan, "Pemanfaatan Algoritma Adasyn Dan Support Vector Machine Dalam Meningkatkan Akurasi Prediksi Kanker Paru-Paru," Jurnal JATI, vol. 8, no. 5, 2024.

[19] T. Fatima, K. Xia, W. Yang, Q. U. Ain, and P. L. Perera, "Diabetes Prediction Using ADASYN-Based Data Augmentation and CNN-BiGRU Deep Learning Model," *Computers, Materials & Continua*, vol. 0, no. 0, pp. 1–10, 2025, doi: 10.32604/cmc.2025.063686.

[20] I. Gede, I. Sudipa, and M. Darmawiguna, *Buku Ajar Data Mining*. 2024. [Online]. Available: https://www.researchgate.net/publication/377415198

[21] W. T. K. Widyastuti Andriyani, Mochammad Anshori, Dwi Normawati, Risqy Siwi Pradini, Mohamad Zaenudin, Muhammad Iqbal Harisuddin, M. Syauqi Haris, Astuty, Anna Angela Sitinjak, *Matematika Pada Kecerdasan Buatan*. Makasar: CV. Tohar Media, 2024.