Performance Evaluation of Tuned and Untuned Machine Learning Models in Speech Emotion Recognition

Muhammad Hudzaifah Nasrullah[1*], Dede Cahyadi[2], Tilly Raycitra Widya[3], Ewin Suciana[4], Lilik Tiara Giantri[5]

[1,2,3,4,5] *Department of Informatics Engineering, Yarsi Pratama University, Indonesia*

*corr-author: hudzaifah@yarsipratama.ac.id

Abstract - This analysis takes on a comparative review of three distinct machine learning approaches: Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and Random Forest (RF) to ascertain emotional states in verbal communication by utilizing the RAVDESS resource. In this review, we perform a strategy that unites audio feature extraction, model training with or without tweaks to hyperparameters, and evaluation via metrics including accuracy, precision, recall, and F1-score. The assessment shows that, before any refinement, SVM secured the utmost accuracy of 79%, trailed by MLP at 76% and RF at 71%. Following optimization, only SVM exhibited an enhancement, reaching 80%, whereas MLP and RF displayed negligible or no improvement. An examination of the confusion matrix revealed that SVM produced the most uniformly distributed predictions and effectively reduced misclassification errors, particularly within the emotion categories of "calm" and "happy." This investigation offers empirical substantiation of SVM as a robust baseline model for speech emotion recognition in localized settings, while simultaneously providing insights into model optimization and development that could inform future implementations in speech-based human–computer interaction.

Keywords: Emotion recognition; machine learning; gridsearchcv; confusion matrix.

I. INTRODUCTION

As a fundamental component of affective computing, Speech Emotion Recognition (SER) seeks to enhance the interaction between human beings and computational systems by enabling these systems to interpret and respond to emotional indicators. This innovation represents a significant step toward the development of more intuitive and empathetic systems, with applications across diverse domains. In healthcare, SER can help detect stress or depression through vocal tone; in education, it can assess student engagement or frustration in online learning environments; in customer service, it allows virtual agents to adapt to a customer's emotional state; and in social robotics, it supports more natural and contextually appropriate interactions [1]. The capacity to adeptly interpret emotions through vocal signals carries substantial consequences for considerably augmenting user interaction, empowering machines to offer replies that are not only practically relevant but also emotionally and contextually aligned [2]. Although considerable strides have been made in Speech Emotion Recognition (SER), the intricate problem of accurately discerning emotions from vocalizations endures due to the complex and often ambiguous nature of human emotions, heightened by the substantial variation in speech patterns affected by accent, tempo, volume, and situational factors [3, 4]. The intricacy of emotional speech data necessitates the development of advanced methodologies and models to effectively address its subtleties.

Recent studies in the SER field have increasingly concentrated on the advancement, modification, and evaluation of diverse machine learning frameworks to enhance classification precision, with classical models like Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and Random Forest (RF) remaining prevalent in speech emotion classification applications [5]. Each model demonstrates unique characteristics, incorporating specific advantages and disadvantages, with their effectiveness differing significantly according to the particular attributes of the dataset and the resultant acoustic and prosodic features [6]. For instance, SVM is widely known for its robustness in handling high-dimensional data and its ability to achieve high classification accuracy in various emotion recognition scenarios [7]. MLP, as a neural network, is effective in modelling complex non-linear patterns but requires extensive training and is sensitive to weight initialization and architecture design [8]. Meanwhile, Random Forest is an ensemble learning technique that integrates various decision trees. It effectively manages large datasets and

reduces overfitting, resulting in high and consistent classification accuracy [9]. These differences highlight the need for comparative evaluation to determine the most effective model for speech emotion classification.

In Indonesia, comprehensive comparative studies on SER remain limited, particularly those analyzing performance before and after hyperparameter optimization with detailed per-class evaluation [10, 11]. This study aims to address this gap by developing an audio-based emotion classification system using the RAVDESS dataset. The evaluation focuses on accuracy, precision, recall, and F1-score for each emotion category, with additional implementation insights. The findings are expected to contribute to the development of more reliable SER systems and provide guidance for future research.

II. METHOD

This research encompasses four phases: audio data acquisition, preprocessing, model development, and performance assessment as shown in Fig. 1.

Three machine learning models, SVM, MLP, and RF, are evaluated through various metrics, including accuracy, precision, recall, and F1-score, both pre- and post-hyperparameter optimization.

A. Dataset

This research utilizes the RAVDESS dataset from Kaggle. The dataset comprises 1440 speech audio files (16bit, 48kHz wav) produced by 24 professional actors (12 female, 12 male), articulating two lexically matched statements in a neutral North American accent. The conveyed emotions encompass calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each emotion is articulated at two emotional intensity levels (normal, strong), alongside a neutral expression. This dataset is derived from the article "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English." [12].

B. Preprocessing

The dataset preprocessing in this study comprises two stages: filtration through emotion code mapping and feature extraction to transform raw audio signals into numerical formats using three methodologies:

1) MFCCs are essential features for speech recognition due to their effectiveness in capturing important speech signal data [13]. MFCC is derived from the speech power spectrum using Mel-frequency bands and DCT for coefficient extraction. This process involves windowing the speech signal, applying Fourier Transforms, and utilizing a Mel-scaled bandpass filter for weighting [14]. While MFCCs effectively capture speech data and align with hidden Markov model assumptions, they demonstrate diminished performance under increased noise conditions.

2) Chroma STFT is a linear time-frequency analysis method that depicts audio segments through complex coefficients of magnitude and phase [15]. The Chroma STFT method enhances time and frequency resolution for non-stationary signal analysis and exhibits linearity, multi-resolution capabilities, and inversion uniqueness. Nonetheless, its fixed window size imposes limitations on time and frequency resolution in audio signal processing, particularly in time-frequency domain analysis.

3) Mel Spectrogram: Mel Spectrogram is a time-frequency representation that aligns the sound power spectrum with the Mel scale, reflecting human auditory perception of frequencies, thus proving effective for voice-related tasks like speech emotion recognition [16]. This technique is pertinent to audio processing tasks, such as genre classification and event detection, often utilizing deep learning models. The Mel Spectrogram effectively captures intricate features, although audio noise can degrade feature quality, which is mitigated through adaptive methods and data augmentation to enhance robustness and discriminability [17].

C. Split Data

The dataset is partitioned into two segments with a distribution of 75% allocated to training and 25% assigned to testing, culminating in 1,080 training audio samples and 360 testing audio samples. Furthermore, the implementation of a seed value of 42 is essential to ensure the reproducibility and stability of the experimental outcomes.

D. Machine Learning Model

This study utilizes Multilayer Perceptron, Support Vector Machine, and Random Forest as machine learning frameworks. These models are chosen because of the limited dataset, providing a basis for future research.

1) The Multilayer Perceptron: represents a hierarchical feed-forward framework comprising an input layer, multiple hidden layers, and an output layer, in which each layer utilizes non-linear activation functions [18]. Unlike standard perceptions, MLPs can classify non-linearly separable data, enhancing their utility in various domains.

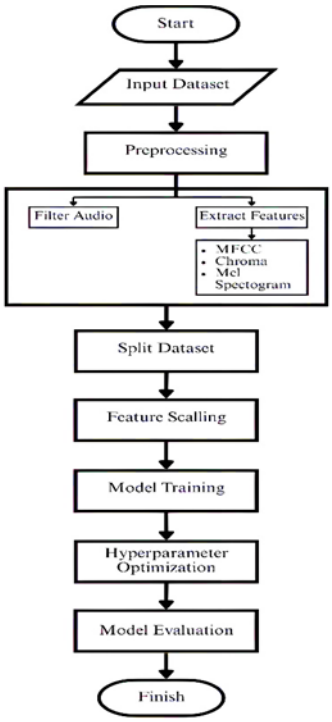

Fig. 1 Research flowchart

2) Support Vector Machine: SVM integrates fundamental machine learning methodologies originally designed for classification, subsequently modified for various domains such as bioinformatics and computer vision [19]. SVM utilizes statistical learning theory and convex optimization to tackle issues related to limited sample sizes, non-linearity, and high dimensionality. SVM is founded on statistical learning theory and convex optimization, allowing it to adeptly address challenges with small sample sizes, non-linearity, and high dimensionality.

3) Random Forest: Random Forest algorithm constitutes a widely utilized ensemble machine learning technique esteemed for its precision and operational efficiency. Its applications extend across various fields, particularly within health analytics for the purpose of disease prediction [20], RF is effective for classifying text, image, and biological data. Its inherent feature selection and interaction capabilities enhance its utility for complex problems and learning tasks.

E. Hyperparameter Tuning

The hyperparameter utilized in this research endeavor is GridSearchCV, which systematically refines machine learning models, including MLP, SVM, and RF, for the purpose of speech emotion classification. This methodology meticulously assesses variations of specified hyperparameters to determine the most advantageous configurations. The specific parameter grids and the number of cross-validation folds utilized for each model are detailed in Table I. Although GridSearchCV markedly improves the accuracy of emotion recognition, its application may demand considerable computational resources, thereby necessitating a careful consideration of the trade-off between enhancements in performance and computational efficiency [21].

Table I delineates the grid parameters and cross-validation folds for each model. SVM involved the evaluation of three C values, two gamma types, and two kernel types. Random Forest utilized parameters such as n_estimators, max_depth, and min_samples_split for optimization. MLP explored diverse configurations of hidden_layer_sizes, activation, alpha, and learning_rate. The tuning methodology employed 5-fold cross-validation to enhance result reliability and reduce dependence on data partitioning.

TABLE I
HYPERPARAMETER GRID AND CROSS-VALIDATION FOLDS

Model	Hyperparameter	Parameter Grid	Number of Cross-Validation Folds
Support Vector Machine	C	[1, 10, 100]	5
	gamma	['scale', 'auto']	5
	kernel	['rbf', 'linear']	5
Random Forest	n_estimators	[50, 100, 200]	5
	max_depth	[None, 10, 20]	5
	min_samples_split	[2, 5, 10]	5
Multi-Layer Perceptron	hidden_layer_sizes	[(256,), (512, 256)]	5
	activation	['relu', 'tanh']	5
	alpha	[0.0001, 0.001]	5
	learning_rate	['constant', 'adaptive']	5

F. Confusion Matrix

The Confusion Matrix serves as a fundamental instrument for the assessment of classification models, offering a comprehensive analysis of predictions in relation to actual outcomes, and methodically categorizing these into True Positives, True Negatives, False Positives, and False Negatives [22]. This matrix is essential for calculating various performance metrics, such as accuracy, precision, recall, and F1 score, which collectively provide a holistic insight into the model's predictive capability [23]. While the confusion matrix provides substantial insights, it is crucial to acknowledge its inherent limitations, particularly in cases of class imbalance or multi-label classification, where it may inadequately represent the nuanced nature of model efficacy or predictive accuracy [24].

III. RESULT AND DISCUSSION

This study examines Support Vector Machine, Multi-Layer Perceptron, and Random Forest models. Prior to model training, the researcher pre-processed the data using feature extraction with MFCC, Chroma STFT and Mel Spectrogram methods. MFCC is used to capture the spectral shape of the human voice based on human auditory perception, Chroma STFT is needed to convert time to frequency signals to produce a local spectral representation and Mel Spectrogram to visualize frequency in the time domain. Preprocessing is done before data splitting to ensure each audio sample is in numerical form.

After splitting the data, the next step is feature scaling with standardscaler to get a normal distribution. This stage is very important to speed up the convergence of gradient descent, improve model accuracy and prevent the dominance of certain features, especially for SVM and MLP models [25].

A. Model Performance Without Optimization

In the initial phase of assessing the voice-based emotion classification system, investigators executed a series of experiments utilizing machine learning algorithms devoid of the implementation of the GridSearchCV hyperparameter optimization methodology. The models utilized in this study include Support Vector Machine, Multi-Layer Perceptron, and Random Forest. This phase seeks to assess the fundamental effectiveness of each model in emotion recognition based on MFCC acoustic features derived from the audio signal.

The MLP architecture consists of two hidden layers featuring 512 and 256 neurons, utilizes ReLU activation, and incorporates an adaptive learning rate. It was trained over 1000 iterations with a batch size of 64. The MLP demonstrated training stability and produced initial predictions for emotion classes without any adjustments to its architecture or learning rate. Thus, it serves as a representative baseline for neural networks in this research. MLP model training results get an accuracy of 75,52%. This result is assisted by the batch size and learning rate to maintain convergence during training, so that the MLP can capture non-linear patterns of voice features, even without model optimization.

The SVM model uses RBF kernel parameters that allow the model to capture complex decisions and the C=10 parameter helps reduce classification errors, resulting in The SVM model achieved the highest performance with 79.17% accuracy. This efficacy is attributed to SVM's capability to manage high-dimensional data and classify non-linear features proficiently. Conversely, the RF model, utilizing 100 decision trees with default parameters, recorded the lowest accuracy at 70.83%. Fig. 2 illustrates a comparison of the accuracy among the three models.

Furthermore, the classification model's performance was assessed using a confusion matrix without optimization. The findings reveal notable disparities in the efficacy of the three machine learning algorithms for emotion classification. SVM exhibited a balanced performance, achieving the highest overall accuracy of 79% across four emotion categories, as indicated by evaluation metrics such as precision, recall, F1-score, and accuracy in Table II. This indicating that the SVM model is suitable to be used as a baseline in voice data-based automatic emotion recognition systems.

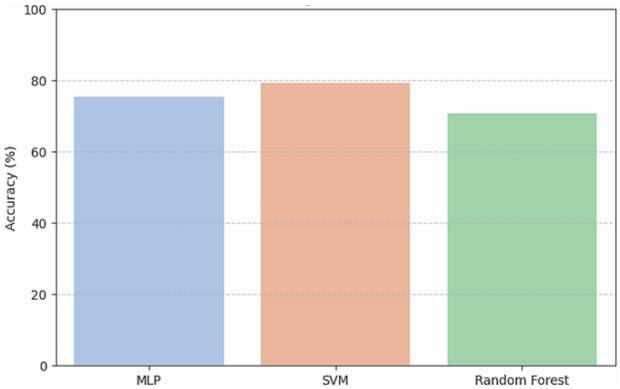

Fig. 2 Comparison of model accuracy before optimization

TABLE II
CONFUSION MATRIX WITH NO OPTIMIZATION

Metrics	Emotion	MLP	SVM	Random Forest
Precision	calm	0.89	0.88	0.74
	disgust	0.79	0.8	0.64
	fearful	0.71	0.77	0.83
	happy	0.65	0.72	0.67
Recall	calm	0.85	0.94	0.96
	disgust	0.73	0.73	0.59
	fearful	0.7	0.72	0.52
	happy	0.75	0.79	0.77
F1-score	calm	0.87	0.91	0.83
	disgust	0.76	0.76	0.61
	fearful	0.7	0.74	0.64
	happy	0.7	0.75	0.72
Accuracy		0.76	0.79	0.71

The MPL model performed quite well close to the SVM model with a total accuracy of 76%. This shows that the MLP model is effective in minimizing false positives and is relevant in audio-based emotion classification applications. The random forest model did not perform consistently across categories but did achieve 96% on the Recall matrix with the calm category and 83% on the Precision matrix with the fearful category. The RF model's default results present a potential avenue for enhancing its accuracy. Visualization of confusion matrix comparisons is crucial to uncover classification biases, misallocations, and model weaknesses across emotional categories, as illustrated in Fig. 3.

B. Model Performance with Optimization

After obtaining the performance results of each algorithm without optimization techniques, the subsequent phase involves enhancing model efficacy via hyperparameter tuning utilizing the GridSearchCV methodology. This method is preferred for its systematic ability to thoroughly explore the parameter landscape through exhaustive search and cross-validation. With refined parameters, each model is anticipated to exhibit enhanced predictive accuracy and stability. The MLP parameters include hidden layer structure, activation function, learning rate, and L2 regulation. For SVM, the optimization parameters comprise penalty attributes, kernel type, and kernel coefficients, whereas Random Forest parameters involve the quantity of trees, maximum tree depth, and the number of features for each split. The optimal generalization scores attained during training are presented in Table III.

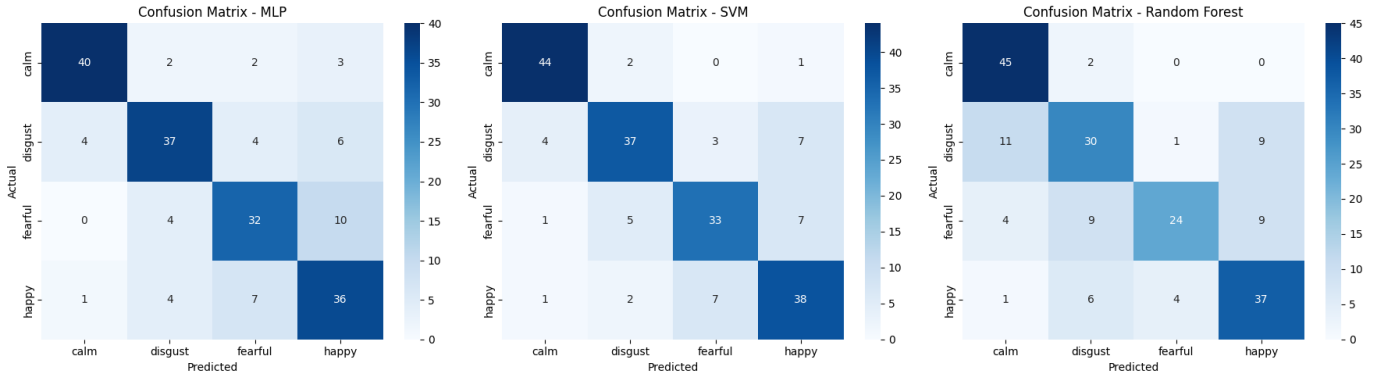

Fig. 3 Comparison of confusion matrix with no optimization

TABLE III
GRIDSEARCHCV OPTIMIZATION RESULT

Model	Best Parameters	Best Accuracy
SVM	{'C': 100, 'gamma': 'scale', 'kernel': 'rbf'}	74.82%
Random Forest	{'max_depth': 10, 'min_samples_split': 2, 'n_estimators': 200}	64.75%
MLP	{'activation': 'relu', 'alpha': 0.0001, 'hidden_layer_sizes': (512, 256), 'learning_rate': 'constant'}	73.44%

The performance of the model after optimization did not improve significantly, especially the accuracy, MLP scored 76.04%, SVM scored 79.69%, and random forest scored 70.31%, as shown in the Fig. 4.

After conducting hyperparameter optimization through GridSearchCV, the models were re-examined for the purpose of enhancing voice emotion classification effectiveness. Evaluation metrics, including precision, recall, F1-score for each emotion category, and overall model accuracy, are presented in Table IV. The comparative visualization of the confusion matrix for each model is illustrated in Fig. 5.

Performance analysis indicates that Support Vector Machine shows outstanding performance with 79.69% accuracy. Multilayer Perceptron showed strong performance at 76.04%. In opposite, Random Forest showed the lowest accuracy at 70.31%. These results show that SVM and MLP are more responsive to hyperparameter modifications for speech emotion classification, unlike Random Forest which experienced a decrease in overall accuracy.

C. Comparison of model performance

After evaluating model performance pre- and post-optimization, the subsequent task involves assessing the models for speech emotion classification with restricted data. This study seeks to identify the model exhibiting superior discriminative and generalization proficiencies on the utilized dataset. The analysis incorporated vital performance measures, featuring accuracy, precision, recall, F1 score, and ROC curve analysis, detailed in Table V.

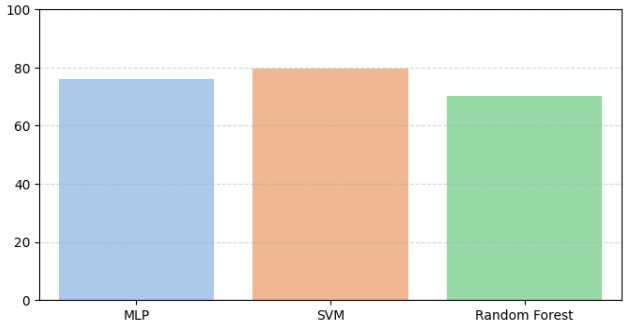

Fig. 4 Comparison of model accuracy after optimization

TABLE IV
CONFUSION MATRIX WITH OPTIMIZATION

Metrics	Emotion	MLP	SVM	Random Forest
Precision	calm	0.89	0.9	0.75
	disgust	0.79	0.78	0.65
	fearful	0.71	0.8	0.76
	happy	0.67	0.71	0.65
Recall	calm	0.85	0.94	0.96
	disgust	0.75	0.82	0.55
	fearful	0.7	0.7	0.57
	happy	0.75	0.73	0.75
F1-score	calm	0.87	0.92	0.84
	disgust	0.77	0.8	0.6
	fearful	0.7	0.74	0.65
	happy	0.71	0.72	0.7
Accuracy		0.76	0.8	0.7

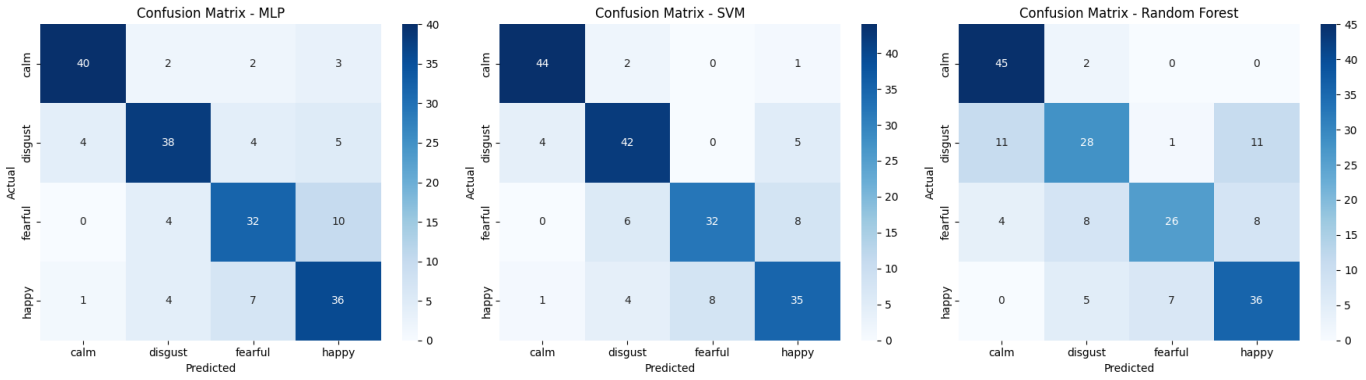

Fig. 5 Comparison of confusion matrix with optimization

TABLE V
MODEL COMPARISON REPORT

Performance Metric	Model	Before Optimization	After Optimization
Accuracy	MLP	0.76	0.76
	SVM	0.79	0.8
	Random Forest	0.71	0.7
Precision	MLP	0.76	0.76
	SVM	0.79	0.8
	Random Forest	0.72	0.71
Recall	MLP	0.76	0.76
	SVM	0.79	0.8
	Random Forest	0.71	0.71
F1-score	MLP	0.76	0.76
	SVM	0.79	0.8
	Random Forest	0.7	0.7

The findings in Table V indicate marginal enhancement in SVM model performance with GridSearchCV optimization, while the MLP model remains unchanged, and the random forest model exhibits a minor decline across various metrics.

The limited efficacy of hyperparameter optimization is linked to various intrinsic factors of the problem and dataset. First, the relatively small size of the RAVDESS dataset restricts performance enhancements through extensive tuning. A larger dataset would provide models with greater learning opportunities from diverse patterns, potentially yielding more notable improvements post-optimization. Second, the models' initial strong performance implies that the default parameters were likely near optimal for this task. Lastly, the specified parameter grid may not encompass configurations capable of producing significant performance advancements. This suggests that for this dataset, the models may be approaching their performance ceiling, necessitating more intricate architectures or feature engineering for further improvements.

The following step entails the evaluation of the multiclass ROC curve for each algorithm, as represented in Fig. 6. The ROC curve serves to illustrate the model's capacity to differentiate among classes across a spectrum of classification thresholds. The macro-AUC values encapsulate the ROC curve, reflecting the model's overall discriminative efficacy. As illustrated by the ROC curves pre- and post-optimization, the SVM model consistently achieves the highest macro-AUC score of 0.94 (Fig. 6), both prior to and after optimization. This underscores the SVM's superior proficiency in accurately classifying emotions across all classes in comparison to MLP and Random Forest.

The MLP model reflects a notable discriminative capability, demonstrated by a macro-AUC score of 0.94 before optimization, with a minimal reduction to 0.93 after optimization. Conversely, the Random Forest model demonstrates the most inferior performance in this context, presenting macro-AUC values of 0.89 and 0.90 prior to and after optimization, respectively. This marginal alteration in macro-AUC metrics substantiates the conclusion that hyperparameter optimization exerted a negligible influence on the models' overall proficiency in class differentiation, a finding that aligns with the observed stability in accuracy outcomes.

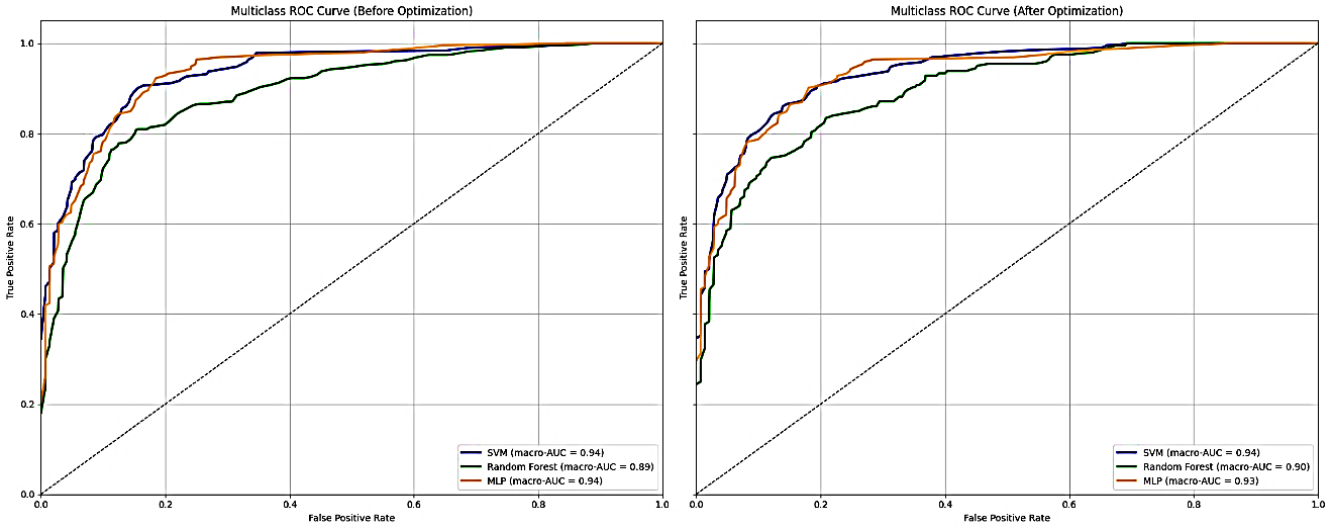

Fig. 6 Comparison of multiclass ROC curve

IV. CONCLUSION

In this study, the authors examine the difficulties associated with emotion discrimination in human vocalizations. A study was performed to gauge the effectiveness of three machine learning frameworks: Support Vector Machine, Multi-Layer Perceptron, and Random Forest, focused on identifying emotions in voice, examining cases with and without hyperparameter optimization via GridSearchCV. Prior to any enhancements, SVM showed the top accuracy standing at 79%, while MLP had 76%, and RF came in at 71%. Following optimization, SVM demonstrated a significant improvement to 80%, alongside enhancements in precision, recall, and F1 score, whereas MLP maintained 76% and RF declined to approximately 70%. The confusion matrix showed that SVM had the most balanced correct predictions and minimal misclassification, especially in the "calm" and "happy" categories. MLP and RF struggled with distinguishing emotions with similar sound patterns, especially for the "fear" and "disgust" classes. The findings suggest future research should improve hyperparameter optimization strategies and explore feature enhancement techniques, including data augmentation and advanced deep learning models. Future research could explore ensemble or hybrid approaches to enhance model performance and reliability.

ACKNOWLEDGEMENT

The scholar recognizes the essential guidance and financial backing offered throughout the duration of the research. Recognition is hereby given to the Rector of Yarsi Pratama University for their crucial role in aiding the research.

REFERENCES

[1] L.-L. Guo, L.-B. Wang, J.-W. Dang, and S.-F. Ding, "Research Progress of Discrete Speech Emotion Recognition," *Ruan Jian Xue Bao*, vol. 35, no. 12, pp. 5487–5508, 2024, doi: 10.13328/j.cnki.jos.007232.

[2] M. O. Oyediran, O. S. Ojo, S. Bharany, A. E. Adeniyi, A. L. Imoize, Y. Farhaoui, and J. B. Awotunde, "Speech emotion recognition using yet another mobile Network tool," in Proc. Int. Conf. Artificial Intelligence and Smart Environment, Cham, Switzerland: Springer International Publishing, 2022, pp. 729-739, doi: 10.1007/978-3-031-26254-8_106.

[3] A. S. Nasim, R. H. Chowdory, A. Dey, and A. Das, "Recognizing speech emotion based on acoustic features using machine learning," in Proc. 2021 Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS), IEEE, 2021, pp. 1–7, doi: 10.1109/ICACSIS53237.2021.9631319.

[4] Y. Li, "Enhancing speech emotion recognition for real-world applications via ASR integration," in Proc. 2023 11th Int. Conf. Affective Comput. Intell. Interact. Workshops Demos (ACIIW), IEEE, 2023, pp. 1–5, doi: 10.1109/ACIIW59127.2023.10388136.

[5] A. Vyakaranam, B. Ramayah, and T. Maul, "Preliminary Study: Speech Emotion Recognition in Online Teaching from the Perspective of Educators Especially Late Deafened," in *Proc. 2024 2nd Int. Conf. Softw. Eng. Inf. Technol. (ICoSEIT)*, 2024, pp. 216–221, doi: 10.1109/ICoSEIT60086.2024.10497503.

[6] T. Rathi and M. Tripathy, "Analyzing the influence of different speech data corpora and speech features on speech emotion recognition: A review," Speech Commun., vol. 162, pp. 103102, 2024, doi: 10.1016/j.specom.2024.103102.

[7] A. R. Lakshminarayanan, I. S. R. Balaji, S. T. Hussain, V. Jayaraman, and C. S. Anwar, "Enhancing Speech Emotional Recognition through a Multi-Layer Perceptron Model," in *Proc. 2023 2nd Int. Conf. Trends Electr., Electron. Comput. Eng. (TEECCON)*, 2023, pp. 178–183, doi: 10.1109/TEECCON59234.2023.10335806.

[8] E. Blumentals and A. Salimbajevs, "Emotion recognition in real-world support call center data for Latvian language," in *CEUR Workshop* Proc., vol. 3124, 2022.

[9] A. V. Porco and D. Kang, "Enhancing Emotion Classification Through Speech and Correlated Emotional Sounds via a Variational Auto-Encoder Model with Prosodic Regularization," in *Proc. 2023 IEEE Int. Conf. Comput. Vis. Mach. Intell. (CVMI)*, 2023, doi: 10.1109/CVMI59935.2023.10464855.

[10] S. Mekruksavanich, A. Jitpattanakul, and N. Hnoohom, "Negative Emotion Recognition using Deep Learning for Thai Language," in *Proc. 2020 Jt. Int. Conf. Digit. Arts, Media Technol. (ECTI DAMT NCON)*, 2020, pp. 71–74, doi: 10.1109/ECTIDAMTNCON48261.2020.9090768.

[11] R. Sharma and A. Pradhan, "Implementation of Machine Learning based Optimized Speech Emotion Recognition," in *Proc. 2nd Int. Conf. Autom., Comput. Renew. Syst. (ICACRS)*, 2023, pp. 1090–1095, doi: 10.1109/ICACRS58579.2023.10405195.

[12] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, p. e0196391, 2018, doi: 10.1371/journal.pone.0196391.

[13] S. Cai, Y. Xiao, J. Pan, Q. Zhao, and Y. Yan, "Noise robust feature scheme for automatic speech recognition based on auditory perceptual mechanisms," *IEICE Trans.*

Inf. Syst., vol. E95-D, no. 6, pp. 1610–1618, 2012, doi: 10.1587/transinf.E95.D.1610.

[14] S. S. Hanna, N. Korany, and M. B. Abd-El-Malek, “Speech recognition using Hilbert-Huang transform based features,” in *Proc. 2017 40th Int. Conf. Telecommun. Signal Process. (TSP)*, 2017, pp. 338–341, doi: 10.1109/TSP.2017.8076000.

[15] S. D. Voran, “Why Some Audio Signal Short-Time Fourier Transform Coefficients Have Nonuniform Phase Distributions,” in *Proc. IEEE Int. Conf. Multimed. Expo (ICME)*, 2024, doi: 10.1109/ICME57554.2024.10687591.

[16] F. L. de Mattos, M. E. Pellenz, and A. S. Britto, “Time Distributed Multiview Representation for Speech Emotion Recognition,” in *Lecture Notes in Computer Science*, 2024, pp. 148–162, doi: 10.1007/978-3-031-49018-7_11.

[17] Y. Tan, Z. Wang, K. Qian, Z. Bao, Z. Cao, B. Hu, Y. Yamamoto, and B. W. Schuller, "Amnet: Introducing an adaptive mel-spectrogram end-to-end neural network for heart sound classification," in Proc. 2023 IEEE Int. Conf. E-health Networking, Application & Services (Healthcom), IEEE, 2023, pp. 90-94, doi: 10.1109/Healthcom56612.2023.10472362.

[18] R. Lin, Z. Zhou, S. You, R. Rao, and C. C. J. Kuo, “Geometrical Interpretation and Design of Multilayer Perceptrons,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 2545–2559, 2024, doi: 10.1109/TNNLS.2022.3190364.

[19] M. Jabardi, “Support Vector Machines: Theory, Algorithms, and Applications,” *Infocommunications J.*, vol. 17, no. 1, pp. 66–75, 2025, doi: 10.36244/ICJ.2025.1.8.

[20] S. Kukreti, K. Al-Attabi, R. Chandrashekar, K. P. Rani, A. Badhoutiya, N. S. Boob, and A. Srivastava, "Enhancing Disease Prediction through Random Forests in Healthcare Analytics," in Proc. 2024 7th Int. Conf. Contemporary Computing and Informatics (IC3I), vol. 7, IEEE, 2024, pp. 1693-1699, doi: 10.1109/IC3I61595.2024.10828927.

[21] A. Thakur and S. K. Dhull, “Language-independent hyperparameter optimization-based speech emotion recognition system,” *Int. J. Inf. Technol. Singap.*, vol. 14, no. 7, pp. 3691–3699, 2022, doi: 10.1007/s41870-022-00996-9.

[22] J. Erbani, P.-É. Portier, E. Egyed-Zsigmond, and D. Nurbakova, "Confusion matrices: A unified theory," IEEE Access, vol. 12, pp. 181372–181419, 2024, doi: 10.1109/ACCESS.2024.3507199.

[23] K. J. S. Narayanan and A. Manimaran, “Using Decision Risk and Decision Accuracy Metrics for Decision Making for Remote Sensing and GIS Applications,” in *Lecture Notes in Civil Engineering*, 2024, pp. 125–136, doi: 10.1007/978-981-99-6229-711.

[24] M. Heydarian, T. E. Doyle, and R. Samavi, “MLCM: Multi-Label Confusion Matrix,” *IEEE Access*, vol. 10, pp. 19083–19095, 2022, doi: 10.1109/ACCESS.2022.3151048.

[25] D. A. Tarihoran and H. Santoso, "Comparative Analysis of Machine Learning Algorithms for Groundwater Potability Classification in Jakarta," JUITA: Jurnal Informatika, vol. 13, no. 3, pp. 371–381, Nov. 2025, doi: 10.30595/juita.v13i3.27348.