A Hybrid SMOTE-Ensemble Framework for Robust Speech Emotion Recognition on Imbalanced Data

Eyad Megdadi[1]*, Mariam Aljouhi[2], Zainab Alblooshi[3], Khaled Shaalan[4]
Faculty of Engineering and IT, British University in Dubai, Dubai 345015, United Arab Emirates[1, 3, 4]
Information Security Engineering Technology Department, Abu Dhabi Polytechnic, Abu Dhabi, United Arab Emirates[2]

***Abstract*—Speech Emotion Recognition (SER) has attracted considerable attention in recent years because of its potential applications in healthcare, intelligent assistants, customer service, and human-computer interaction. However, the performance of SER systems is often affected by the imbalance of emotion classes in benchmark datasets, where minority emotions are underrepresented and difficult to classify accurately. This study reproduces the work of Sahu et al. on multimodal Speech Emotion Recognition using the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset. Further, it proposes a Hybrid SMOTE-Ensemble Framework to improve classification performance on imbalanced data. The reproduction stage follows the original methodology by employing handcrafted acoustic features together with TF-IDF textual features under Audio-only, Text-only, and Audio+Text modalities. The enhancement stage introduces two XGBoost classifiers trained independently on the original imbalanced dataset and on a SMOTE-balanced dataset. Their prediction probabilities are combined using three ensemble strategies, namely Simple Averaging, Confidence-Based Selection, and Adaptive Weighting. In addition, the effectiveness of the proposed strategy is further investigated using a LinearSVC classifier. Experimental results demonstrate that the reproduced models achieve results that are highly consistent with those reported in the original study. Furthermore, the proposed Hybrid SMOTE-Ensemble Framework improves the recognition of minority emotion classes while maintaining balanced overall performance. The XGBoost model trained on SMOTE-balanced data achieves higher F1-score and recall than the baseline model, whereas the ensemble approaches provide additional improvements in classification performance. The LinearSVC experiments further confirm that combining oversampling and ensemble learning offers a robust solution to the class imbalance problem in Speech Emotion Recognition.**

***Keywords*—IEMOCAP dataset; Speech Emotion Recognition (SER); class imbalance; SMOTE**

I. Introduction

Emotion Recognition (SER) is an important research area in Artificial Intelligence (AI) that combines signal processing, machine learning, and affective computing. The main goal of SER is to automatically identify the emotional state of a speaker from speech signals. Unlike conventional speech recognition systems, which focus on recognizing spoken words, SER aims to understand the emotions hidden behind speech. This capability enables machines to interact with humans more naturally and intelligently [3].

In recent years, SER has attracted considerable attention because of its wide range of applications. Emotion-aware systems are increasingly used in healthcare, customer service, intelligent virtual assistants, education, and smart transportation systems. For example, intelligent assistants can adapt their responses according to the emotional state of users, while healthcare applications can use speech emotion analysis to monitor mental health conditions such as stress, anxiety, and depression. Similarly, driver monitoring systems can detect emotions such as anger or fatigue to improve road safety [3].

The rapid development of machine learning and deep learning techniques has significantly improved the performance of SER systems. Traditional machine learning approaches rely on handcrafted acoustic features such as pitch, energy, harmonics, pauses, and statistical descriptors. More recently, multimodal approaches that combine speech and textual information have shown promising results because emotional information is often conveyed through both acoustic and linguistic cues [3].

Despite these advances, SER remains a challenging task. One of the main challenges is the imbalance of emotion classes in benchmark datasets. In many real-world datasets, some emotions are represented by a large number of samples, whereas others contain only a limited number of examples. As a result, classification models tend to be biased toward majority classes and often fail to recognize minority emotions accurately. This problem becomes more significant in widely used SER datasets such as the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset, where emotions such as fear and surprise are underrepresented compared with other emotion categories [4].

Several methods have been proposed to address the class imbalance problem in SER. Traditional approaches include random oversampling, undersampling, and cost-sensitive learning. Among these methods, the Synthetic Minority Over-sampling Technique (SMOTE) has gained considerable attention because it generates synthetic samples for minority classes instead of simply duplicating existing instances. By creating new samples through interpolation between neighboring minority samples, SMOTE increases data diversity and reduces the risk of overfitting [6]. Previous studies have shown that SMOTE can improve the performance of classification models on imbalanced datasets and enhance the recognition of minority classes [7,8]. However, SMOTE also has some limitations, including the possibility of generating noisy samples, increasing class overlap, and producing synthetic instances that do not accurately represent the underlying data distribution [9,10].

Ensemble learning has also demonstrated strong performance in many classification problems by combining the

*Corresponding author.

strengths of multiple models. In SER, ensemble approaches can improve classification accuracy and robustness by exploiting the complementary characteristics of different classifiers. Nevertheless, effectively combining ensemble learning with oversampling techniques for imbalanced emotion recognition remains an open research problem. Existing methods often improve minority class recognition at the expense of majority classes or fail to maintain balanced overall performance.

Motivated by these challenges, this study proposes a Hybrid SMOTE-Ensemble Framework for robust Speech Emotion Recognition on imbalanced data. The proposed framework combines models trained on the original imbalanced dataset and on a SMOTE-balanced dataset using different ensemble strategies. The objective is to improve minority emotion recognition while preserving the overall classification performance. The effectiveness of the proposed framework is evaluated using the IEMOCAP dataset and compared with conventional machine learning approaches.

The remainder of this study is organized as follows. Section II presents the background of Speech Emotion Recognition and discusses existing approaches for handling class imbalance. Section III describes the proposed Hybrid SMOTE-Ensemble Framework, including dataset preparation, feature extraction, model development, and ensemble strategies. Section IV presents the experimental setup and evaluation metrics. Section V discusses the experimental results and comparative analysis. Finally, Section VI concludes the study and outlines possible directions for future work.

II. Background

A. Multimodal Speech Emotion Recognition

Speech Emotion Recognition (SER) is an important research field that aims to identify the emotional state of a speaker from speech signals. The objective of SER is to enable machines to understand human emotions and improve human-computer interaction. In recent years, SER has been applied in many areas such as intelligent virtual assistants, healthcare monitoring, customer service, and smart transportation systems [3].

Emotions are usually expressed through multiple channels at the same time. Humans naturally combine information from speech, text, facial expressions, and gestures to understand emotions. Similarly, multimodal SER systems utilize information from different modalities to improve recognition accuracy. Sahu et al. [3] stated that humans are able to resolve ambiguity in emotions because they can efficiently process information from multiple domains, including speech, text, and visual signals. This observation has motivated the development of multimodal SER systems that combine acoustic and textual information to provide a more comprehensive understanding of emotional expressions.

In audio-based SER, acoustic features play an important role in emotion classification. Commonly used features include prosodic characteristics such as pitch, energy, speaking rate, and voice quality. Spectral features such as Mel Frequency Cepstral Coefficients (MFCCs), formants, harmonics, and statistical descriptors are also widely used because they capture important information related to emotional variations in speech [3]. In addition to acoustic information, textual features extracted from speech transcriptions can provide useful semantic information. Traditional text representations include Bag-of-Words and Term Frequency-Inverse Document Frequency (TF-IDF), while more recent approaches employ word embeddings and pre-trained language models.

Among the available SER datasets, the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset is one of the most widely used benchmarks. The dataset contains approximately 12 hours of multimodal recordings collected from ten actors participating in scripted and improvised dialogues [4]. Each utterance is annotated by multiple annotators according to different emotional categories, making the dataset suitable for evaluating speech emotion recognition methods. However, similar to many real-world datasets, IEMOCAP exhibits an imbalanced distribution of emotion classes, where some emotions are represented by a large number of samples while others contain relatively few examples. This imbalance remains one of the major challenges affecting the performance and robustness of SER systems.

B. Class Imbalance and Oversampling Techniques

Class imbalance is one of the major challenges in Speech Emotion Recognition (SER). In many emotional speech datasets, some emotions are represented by a large number of samples, while others contain only a limited number of examples. As a result, machine learning models tend to favor majority classes and often produce poor recognition performance for minority emotions. This issue becomes more serious when the evaluation focuses only on overall accuracy, as a high accuracy score may hide weak performance on underrepresented classes.

The IEMOCAP dataset is a clear example of this problem. The distribution of emotion classes in IEMOCAP is not balanced, with emotions such as fear and surprise containing fewer samples than emotions such as sadness and neutrality. Previous studies reported that this imbalance negatively affects the performance of emotion classification models, particularly for minority emotions that are difficult to learn because of the limited training data available [3]. Therefore, addressing class imbalance has become an important research topic in SER [11].

Traditional methods for handling imbalanced data include random undersampling and random oversampling. Undersampling reduces the number of majority class samples to balance the dataset; however, useful information may be lost during this process. On the other hand, random oversampling duplicates minority class samples to increase their representation. Although this method is simple and easy to implement, it does not introduce new information and may increase the risk of overfitting because the same samples are repeatedly used during training [6].

Among the available oversampling techniques, the Synthetic Minority Over-sampling Technique (SMOTE) is one of the most widely used methods for imbalanced classification problems. SMOTE generates new synthetic samples for minority classes instead of simply duplicating existing instances. The method creates new samples by interpolating between a minority sample and its nearest neighbors in the feature space. This process increases the diversity of minority samples and helps classifiers

learn more representative decision boundaries [6]. Previous studies have demonstrated that SMOTE can improve classification accuracy, precision, recall, and F1-score when dealing with imbalanced datasets [7,8].

Several extensions of SMOTE have also been proposed to improve the quality of synthetic samples. SMOTEBoost combines SMOTE with boosting algorithms to improve the classification of minority classes by generating synthetic samples during the boosting process. Selective Interpolation SMOTE (SISMOTE) focuses on generating samples in safe regions of the feature space to reduce the influence of noisy samples and overlapping classes [2]. Another variant, FC-SMOTE, utilizes fuzzy clustering to identify suitable oversampling regions and assign sampling weights according to data density, which helps address within-class imbalance and improve minority class representation [12].

In recent years, more advanced data augmentation techniques have been introduced for SER. Generative Adversarial Networks (GANs) have been employed to generate realistic synthetic emotional speech samples by learning the underlying data distribution [13]. Other approaches combine acoustic and textual augmentation techniques to preserve emotional consistency across different modalities [14]. Although these methods have shown promising results, they often require complex training procedures and considerable computational resources.

Despite the success of these oversampling and augmentation methods, several challenges remain. Synthetic samples may not always accurately represent the original data distribution, which can lead to class overlap or the introduction of noisy samples. Furthermore, improving minority class recognition may sometimes reduce the performance of majority classes, making it difficult to achieve balanced overall performance [9,10]. Therefore, combining oversampling techniques with ensemble learning has emerged as a promising direction for improving the robustness and effectiveness of Speech Emotion Recognition systems on imbalanced datasets.

C. Ensemble Learning for Imbalanced SER

Ensemble learning is a machine learning technique that combines the predictions of multiple classifiers to achieve better performance than a single model. The main idea is that different classifiers may learn different characteristics from the data, and combining their outputs can improve classification accuracy, robustness, and generalization ability [16]. Ensemble methods have been successfully applied to many classification problems, particularly when dealing with noisy and imbalanced datasets [14].

In Speech Emotion Recognition (SER), ensemble learning has received increasing attention because emotional speech data exhibit large variations in speaking style, recording environments, and emotion distributions. Individual classifiers may perform well on certain emotion classes but perform poorly on others. Therefore, combining multiple classifiers can help reduce the weaknesses of individual models and provide more stable performance across different emotion categories [17-20].

Several ensemble techniques have been investigated in SER research. Voting-based ensembles combine the predictions of different classifiers using majority voting or probability averaging. Boosting algorithms, such as AdaBoost and Gradient Boosting, iteratively train weak learners by focusing on previously misclassified samples to improve the overall prediction performance [21-25]. Bagging approaches, such as Random Forest, create multiple classifiers from different subsets of the training data and aggregate their predictions to reduce variance and improve stability. More recently, hybrid ensemble frameworks have been proposed to combine different classifiers and feature representations for emotion recognition tasks [26-30].

Ensemble learning is particularly useful for imbalanced datasets. Models trained on the original dataset usually preserve the natural distribution of the data and often achieve good performance on majority classes. In contrast, models trained on balanced datasets generated using oversampling techniques, such as SMOTE, generally provide better recognition of minority classes [5]. Combining these models enables the ensemble to exploit the advantages of both learning strategies and can lead to more balanced classification performance [31-35].

However, designing an effective ensemble framework remains a challenging task. The performance of an ensemble depends not only on the quality of the individual classifiers but also on the strategy used to combine their predictions [35-40]. Simple averaging may not always produce the best results, while confidence-based and adaptive weighting methods may behave differently depending on the class distribution and dataset characteristics [15]. Furthermore, the combination of ensemble learning and oversampling techniques in SER has not been extensively investigated, especially for highly imbalanced emotion datasets such as IEMOCAP [4].

Motivated by these observations, this study proposes a Hybrid SMOTE-Ensemble Framework that combines models trained on the original imbalanced dataset and on a SMOTE-balanced dataset using different ensemble strategies. The proposed framework aims to improve the recognition of minority emotions while maintaining robust overall classification performance.

III. Proposed Hybrid SMOTE-Ensemble Framework

This section presents the proposed Hybrid SMOTE-Ensemble Framework for Speech Emotion Recognition on imbalanced data. The framework consists of three main stages: dataset preparation, feature extraction, and hybrid ensemble learning. The IEMOCAP dataset is used to evaluate the effectiveness of the proposed framework because it is one of the most widely used benchmarks for Speech Emotion Recognition and contains a naturally imbalanced distribution of emotion classes [4].

The proposed framework first prepares the emotional speech data and performs emotion label mapping to obtain the final emotion categories. Acoustic and textual features are then extracted from speech recordings and transcriptions to construct multimodal feature representations. To address the class imbalance problem, two classifiers are trained independently: one using the original imbalanced dataset and another using a balanced dataset generated using the Synthetic Minority Over-

sampling Technique (SMOTE). Finally, different ensemble strategies are employed to combine the predictions of the two models and improve the recognition of minority emotions while maintaining robust overall classification performance.

A. Dataset and Preprocessing

The proposed framework was evaluated using the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset, which is one of the most widely used benchmark datasets for Speech Emotion Recognition (SER) [4]. The dataset contains approximately 12 hours of multimodal recordings collected from five sessions of dyadic interactions involving ten actors. Each session includes both scripted and improvised conversations and provides audio recordings, video streams, facial motion capture data, and textual transcriptions. The utterances are annotated by multiple evaluators according to different emotion categories, making the dataset suitable for emotion recognition studies.

The original IEMOCAP dataset contains several emotion labels, including anger, excitement, fear, sadness, surprise, frustration, happiness, disappointment, and neutrality. In this study, the emotion labels were processed to obtain six final emotion categories following the data preparation strategy adopted in previous studies [12]. The emotions "happy" and "excited" were merged into a single class because of their close semantic and expressive characteristics, and because the number of "happy" samples alone was relatively small. The class labeled as "others" was excluded because it represents vague emotional expressions that are difficult to classify consistently.

The emotion label "frustration" does not appear in the final six emotion classes reported by Sahu [12]. Therefore, utterances labeled as "frustration" were assigned to the "angry" category because both emotions share similar emotional characteristics and are often closely related in Speech Emotion Recognition studies. Other ambiguous labels, such as "disappointed", were excluded from the final dataset unless they were explicitly assigned to one of the selected emotion classes.

After label processing, the final dataset consisted of six emotion classes: Angry, Happy, Sad, Fear, Surprise, and Neutral. The corresponding class distribution was adopted from the dataset preparation procedure reported in [12], resulting in 860 samples for Angry, 1309 samples for Happy, 2327 samples for Sad, 1007 samples for Fear, 949 samples for Surprise, and 1385 samples for Neutral, with a total of 7837 utterances. The distribution of these classes remains imbalanced, where some emotions are represented by considerably fewer samples than others.

To evaluate the proposed framework, the dataset was randomly divided into training and testing sets using an 80:20 ratio. The same data split was used throughout all experiments to ensure a fair comparison among different models and ensemble strategies [12]. The original imbalanced training set was preserved for training the baseline model, while an additional balanced training set was generated using the Synthetic Minority Over-sampling Technique (SMOTE) for the proposed hybrid ensemble framework.

B. Feature Extraction

Feature extraction is an important stage in Speech Emotion Recognition because the quality of extracted features directly affects the performance of classification models. In this study, both acoustic and textual features were utilized to construct a multimodal representation of emotional speech following the feature extraction strategy described in [12].

For the audio modality, handcrafted acoustic features were extracted from each speech utterance. These features include pitch, harmonics, energy, pause-related information, and statistical descriptors that capture the emotional characteristics of speech signals. Prosodic and spectral features are widely used in SER because they provide useful information related to speaking style, emotional intensity, and vocal characteristics [3].

For the textual modality, the speech transcriptions provided in the IEMOCAP dataset were transformed into numerical feature vectors using the Term Frequency-Inverse Document Frequency (TF-IDF) technique. TF-IDF was employed to capture the importance of words according to their frequency in a document and their occurrence across the entire corpus. The extracted text features provide complementary linguistic information that can help distinguish different emotional states.

The extracted audio and text features were then combined to form a multimodal feature representation. This combination allows the framework to exploit complementary information from both modalities, since emotions are often reflected not only in acoustic patterns but also in the semantic content of speech.

It should be noted that some implementation details related to audio feature extraction, such as frame size, hop length, and filter configurations, were not explicitly reported in the original framework [12]. Therefore, commonly adopted settings and library default parameters were used during feature extraction while maintaining consistency across all experiments. The same feature extraction procedure was applied to both the original and SMOTE-balanced datasets to ensure a fair comparison among the evaluated models.

C. Model Architectures and Training

To address the class imbalance problem in Speech Emotion Recognition, two XGBoost classifiers were trained independently using different training datasets. The first classifier was trained using the original imbalanced training set, while the second classifier was trained using a balanced dataset generated using the Synthetic Minority Over-sampling Technique (SMOTE). This strategy allows the framework to benefit from the strengths of both learning schemes, where the original model preserves the natural data distribution and the SMOTE-based model improves the recognition of minority emotion classes.

XGBoost was selected as the base classifier because of its strong classification performance, robustness, and ability to handle high-dimensional feature spaces efficiently. XGBoost is an ensemble learning algorithm based on gradient boosting decision trees, where multiple weak learners are sequentially combined to produce a strong predictive model. The algorithm has demonstrated competitive performance in various classification tasks and has been successfully applied in Speech Emotion Recognition applications [1].

For the original model, the extracted multimodal features were directly used to train the XGBoost classifier using the imbalanced training dataset. In parallel, SMOTE was applied to the training set to generate synthetic samples for minority emotion classes and create a balanced training dataset. The same XGBoost architecture and training settings were then applied to the SMOTE-balanced dataset to ensure a fair comparison between the two models.

After training the individual classifiers, their predictions were combined using a hybrid ensemble strategy. Three ensemble methods were investigated in this study: Simple Averaging, Confidence-Based Selection, and Adaptive Weighting. In the Simple Averaging method, the prediction probabilities of the two models were averaged, and the emotion class with the highest probability was selected as the final prediction. In the Confidence-Based Selection method, the prediction of the model with the higher confidence score was chosen. The Adaptive Weighting approach assigns different weights to the two classifiers according to their performance, allowing the ensemble to balance the contributions of the original and SMOTE-based models.

The proposed Hybrid SMOTE-Ensemble Framework combines the advantages of the original and balanced datasets while reducing the limitations of using either approach independently. The overall architecture of the proposed framework is illustrated in Fig. 1.

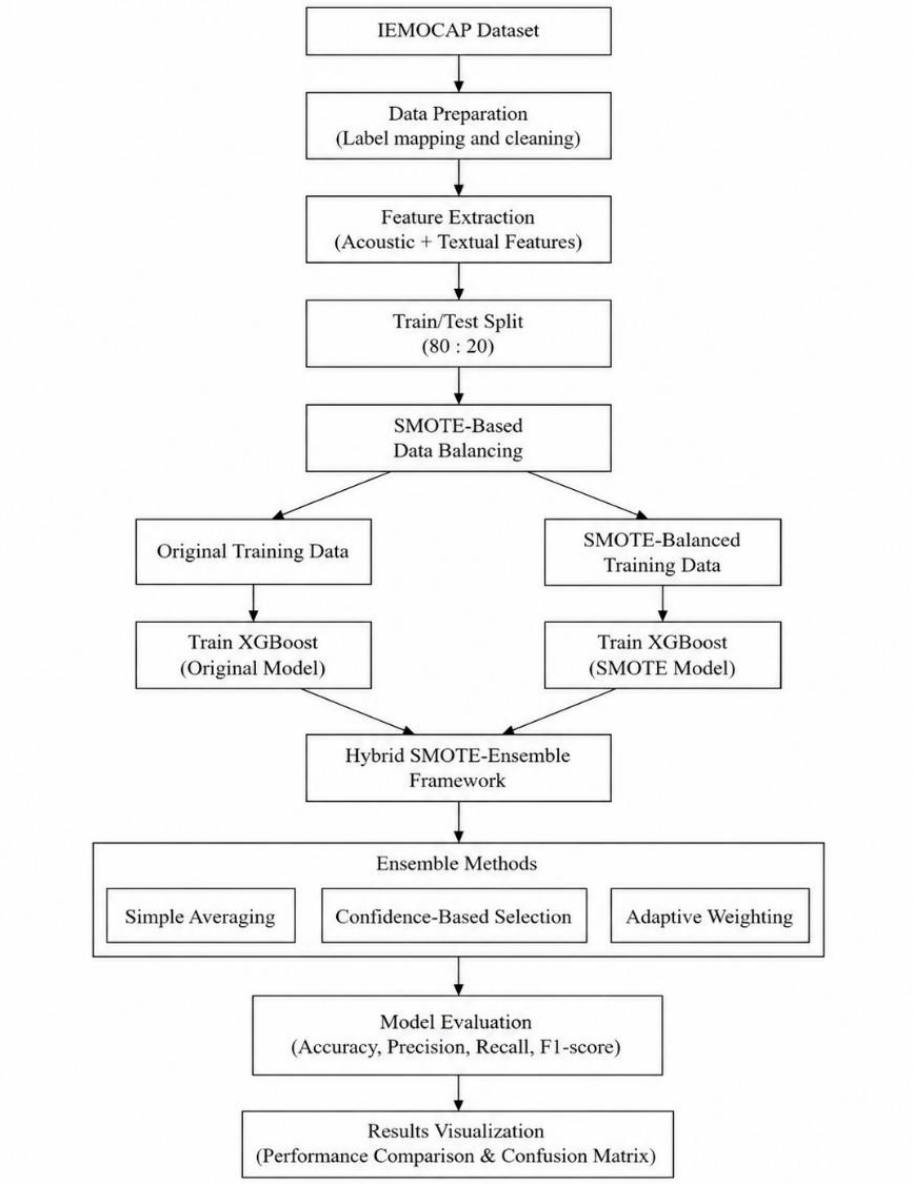

Fig. 1. Overall architecture of the proposed hybrid SMOTE-ensemble framework for speech emotion recognition.

IV. Experimental Setup and Enhancement Strategy

A. Motivation for Enhancement

The IEMOCAP dataset exhibits an imbalanced distribution of emotion classes, where some emotions are represented by a large number of samples while others contain relatively few examples. This imbalance can bias classification models toward majority classes and reduce their ability to recognize minority emotions accurately. In the original study, the imbalance problem was addressed using simple random upsampling for the "fear" and "surprise" classes, while the "happy" and "excited" classes were merged into a single category [12]. Although these strategies help reduce the imbalance problem, they have certain limitations.

Random oversampling increases the number of minority class samples by duplicating existing instances. While this approach is simple to implement, it does not introduce new information into the training set and may increase the risk of overfitting. As a result, the classifier may perform well on the duplicated training samples but fail to generalize effectively to unseen data. Furthermore, merging emotion classes simplifies the classification problem but may reduce the emotional diversity of the dataset and limit the model's ability to distinguish between similar emotions.

To overcome these limitations, this study employs the Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples for minority emotion classes. Unlike random oversampling, SMOTE creates new samples by interpolating between neighboring minority instances, which increases the diversity of the training data and provides better representation of underrepresented classes [5]. This property makes SMOTE a suitable choice for Speech Emotion Recognition tasks involving imbalanced datasets.

In addition to SMOTE, this study introduces a hybrid ensemble strategy that combines models trained on the original imbalanced dataset and on the SMOTE-balanced dataset. The motivation behind this approach is that the original model preserves the natural distribution of the data and performs well on majority classes, whereas the SMOTE-based model is expected to improve the recognition of minority emotions. By combining the predictions of both models, the proposed framework aims to achieve more balanced and robust emotion recognition performance across all emotion classes.

B. Implementing the Hybrid XGBoost Ensemble with SMOTE

To reduce the limitations of simple upsampling and improve model performance, this study uses a hybrid XGBoost ensemble strategy with SMOTE. The method combines the predictions of two separately trained XGBoost models: one trained on the original imbalanced data and another trained on the SMOTE-balanced data.

The first XGBoost model was trained on the original training dataset. This model keeps the natural class distribution and is expected to perform better on the majority emotion classes. The second XGBoost model was trained on a balanced version of the training dataset after applying SMOTE. This model is expected to improve the recognition of minority emotion classes because it is trained with a more balanced class distribution.

SMOTE generates synthetic samples for minority classes by creating new points between existing minority samples and their nearest neighbors in the feature space [5]. In this study, SMOTE was applied using random_state = 42 for reproducibility, and the default k_neighbors = 5 was used. Fig. 2 shows the class distribution before and after applying SMOTE, where all classes were balanced to the same number of samples.

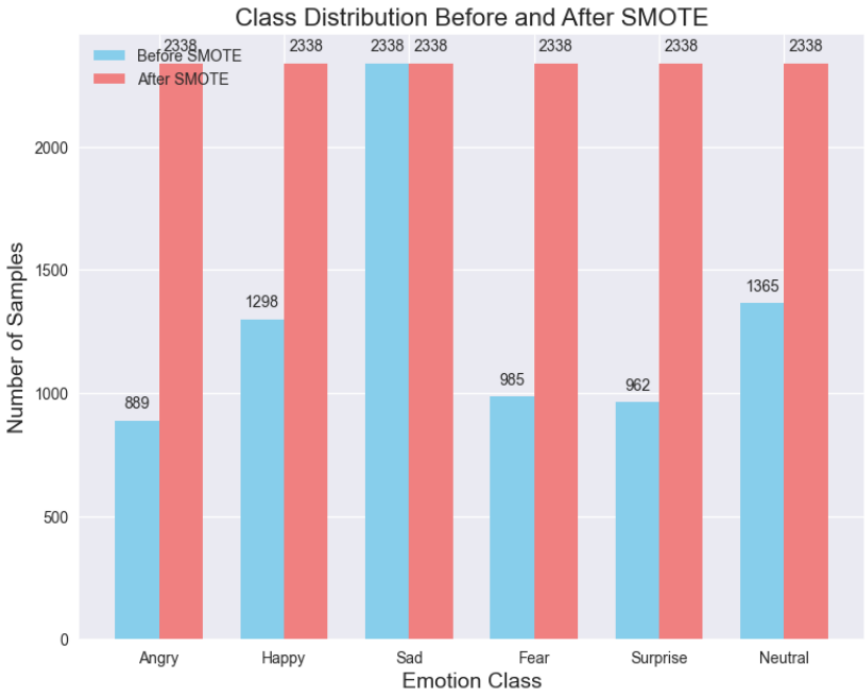

Fig. 2. Class distribution before and after applying SMOTE.

Both XGBoost models were trained using the same parameters. The selected parameters were max_depth = 7, learning_rate = 0.008, objective = multi:softprob, n_estimators = 600, subsample = 0.8, num_class = 6, and booster = gbtree.

The prediction probabilities from the two XGBoost models were combined using three ensemble strategies. The first strategy was Simple Averaging, where the probabilities from the original model and the SMOTE-based model were averaged, and the class with the highest average probability was selected. The second strategy was Confidence-Based Selection, where the final prediction was taken from the model with the highest confidence score. Confidence was calculated as the highest probability assigned to any class by each model.

The third strategy was Adaptive Weighted Ensemble. In this method, different weights were assigned to the original and SMOTE-based models according to the class distribution in the training data. For minority classes, defined as classes with less than 15% of the training samples, the original model was assigned a weight of 0.3 and the SMOTE model was assigned a weight of 0.7. For majority classes, defined as classes with more than 25% of the training samples, the original model was assigned a weight of 0.7 and the SMOTE model was assigned a weight of 0.3. For the remaining classes, both models were assigned equal weights of 0.5.

The purpose of this hybrid ensemble is to combine the strengths of both models. The original XGBoost model can preserve performance on majority classes, while the SMOTE-based XGBoost model can improve the recognition of minority classes. By combining both models, the proposed framework aims to improve the overall Speech Emotion Recognition performance on imbalanced data.

C. Implementation of the Hybrid Ensemble

The proposed Hybrid SMOTE-Ensemble Framework was implemented in Python using the XGBoost and imbalanced-learn libraries. Two XGBoost classifiers were trained independently: one using the original imbalanced training dataset and the other using the SMOTE-balanced dataset. The Audio+Text modality was adopted using the handcrafted acoustic features and TF-IDF text features described in the previous sections. The dataset was divided into 80% for training and 20% for testing, where the testing set remained unchanged throughout all experiments. The prediction probabilities generated by the two classifiers were combined using the three ensemble strategies discussed in Section IV-B, namely Simple Averaging, Confidence-Based Selection, and Adaptive Weighting. The effectiveness of the proposed framework was evaluated using Accuracy, F1-score, Precision, Recall, and confusion matrices.

Fig. 3 presents the performance comparison of the Hybrid XGBoost framework using five different techniques: XGB Original, XGB SMOTE, Simple Average Ensemble, Confidence-Based Ensemble, and Adaptive Weighted Ensemble. The comparison is based on Accuracy, F1-score, Precision, and Recall. The results show that the SMOTE-based model improves both recall and F1-score compared with the original XGBoost model. Furthermore, the ensemble approaches provide additional improvements in overall performance. Among the ensemble methods, the Adaptive Weighted Ensemble demonstrates a balanced improvement across all evaluation metrics, while the Simple Average Ensemble achieves competitive performance with a simpler integration strategy.

Fig. 4 illustrates the confusion matrix of the XGBoost model trained on the original imbalanced dataset. The model classifies emotions into six categories: angry (ang), happy (hap), neutral (neu), sad (sad), frustrated (fru), and excited (exc). The results indicate that the model performs best on the neutral class, which records the highest number of correctly classified samples. However, some confusion can be observed between emotionally similar classes, particularly between happy and neutral emotions. Although misclassifications occur across several categories, the model maintains acceptable classification performance overall. The darker values along the diagonal indicate higher numbers of correct predictions.

Fig. 3. Hybrid XGBoost model performance metrics with different applied techniques.

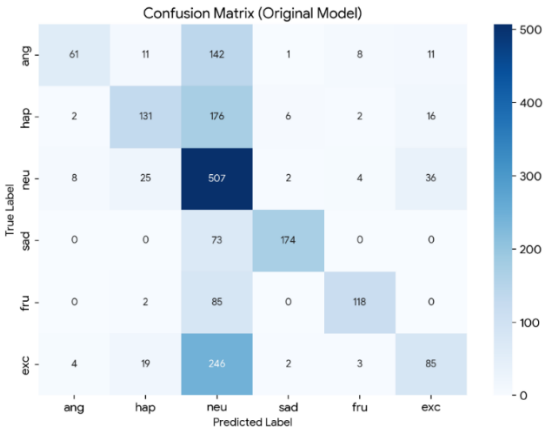

Fig. 4. Confusion matrix of the XGBoost original model.

Fig. 5 presents the confusion matrix of the Simple Average Ensemble. Similar to the original model, the ensemble classifies six emotion categories: angry, happy, neutral, sad, frustrated, and excited. The diagonal elements represent correctly classified samples, while the off-diagonal elements correspond to misclassified instances. The neutral emotion achieves the highest number of correct classifications, whereas confusion remains noticeable between happy and neutral emotions. Overall, the ensemble demonstrates improved consistency across emotion categories and provides more balanced classification performance than the original model.

Fig. 6 shows the confusion matrix of the XGBoost model trained on the SMOTE-balanced dataset. The application of SMOTE improves the representation of minority emotion classes and enhances their recognition capability. The model performs well in recognizing neutral, happy, and excited emotions, while some confusion remains between neutral and sad emotions. The results indicate that balancing the training data contributes positively to the classification of underrepresented classes and improves the overall robustness of the model.

Fig. 7 illustrates the recall values achieved by the Original XGBoost model, the SMOTE-based XGBoost model, and the Simple Average Ensemble across the six emotion classes. The Original model achieves the highest recall for the neutral class, while the SMOTE-based model improves the recall of several minority classes. The Simple Average Ensemble provides more consistent recall values across most classes and achieves a better balance between majority and minority emotions. Nevertheless, the angry class remains the most challenging category for all models, indicating that further improvements in feature representation or data augmentation may still be beneficial.

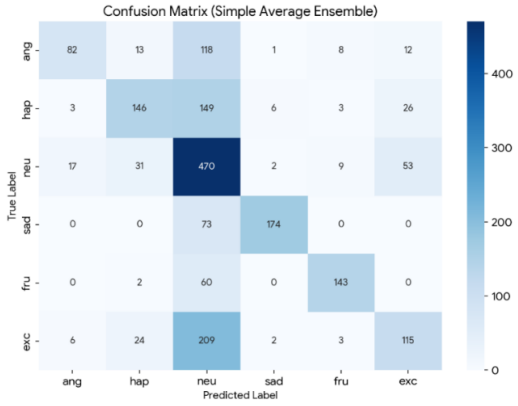

Fig. 5. Confusion matrix of the simple average ensemble technique.

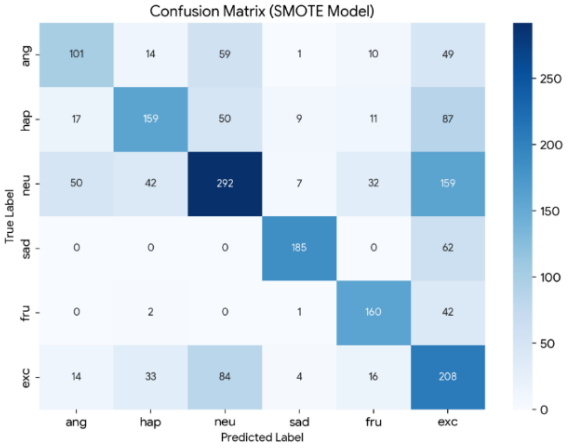

Fig. 6. Confusion matrix of the SMOTE model technique.

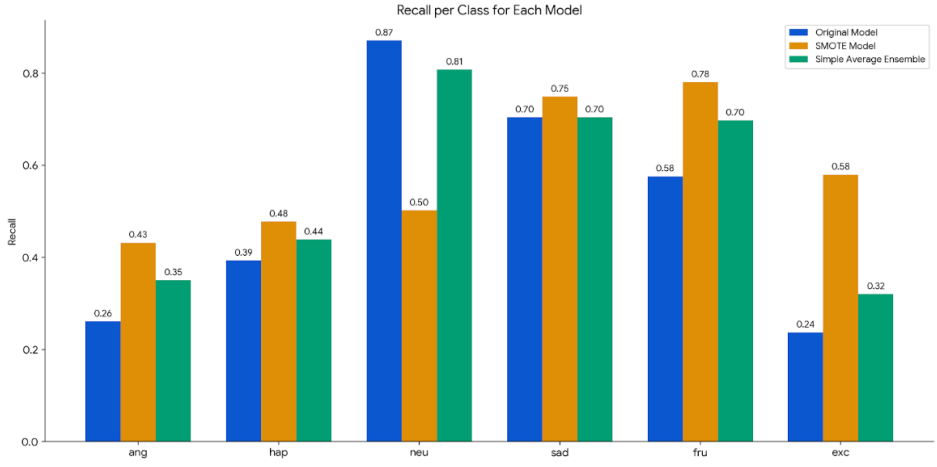

Fig. 7. Hybrid XGBoost model recall per class for each model.

V. Additional Evaluation Using LinearSVC

To further investigate the effectiveness of the proposed Hybrid SMOTE-Ensemble Framework, an additional evaluation was conducted using a Linear Support Vector Classifier (LinearSVC). The objective of this experiment was to examine whether the advantages of data balancing and ensemble learning can be generalized to a different classifier architecture. Similar to the XGBoost experiments, the main challenge lies in the imbalanced distribution of emotion classes, where minority emotions are often underrepresented and difficult to classify accurately.

Two enhancement approaches were investigated. The first approach applied SMOTE directly to the entire training set before training a single LinearSVC classifier. The second approach employed a Hybrid Weighted Ensemble, where two LinearSVC models were trained independently: one using the original imbalanced dataset and the other using the SMOTE-balanced dataset. The predictions of both models were then combined using adaptive weights that assign greater importance to the SMOTE model for minority classes and greater importance to the original model for majority classes.

A. Performance Comparison

The results demonstrate that applying SMOTE changes the behaviour of the LinearSVC classifier. Although the overall accuracy remained close to the baseline model, SMOTE improved the model's ability to recognize minority emotion classes by increasing recall. This improvement, however, was accompanied by a slight reduction in precision and F1-score for some majority classes, which is a common trade-off when balancing highly imbalanced datasets.

The Hybrid Weighted Ensemble achieved the best overall performance among the evaluated LinearSVC models. The ensemble obtained an accuracy score of 0.629, outperforming both the baseline and the SMOTE-only models. This result indicates that combining the strengths of the original and SMOTE-based classifiers leads to a more balanced and robust classification system. In particular, the ensemble achieved perfect recall for the Fear class, demonstrating its effectiveness in recognizing minority emotions.

Fig. 8 illustrates the effect of SMOTE on the class distribution and presents the performance metrics of the LinearSVC models. The results show that the Hybrid Ensemble achieves the highest accuracy, F1-score, and recall, while maintaining competitive precision. In contrast, the original LinearSVC model achieves the highest precision but exhibits lower recall, especially for minority classes. The SMOTE-based model improves recall but sacrifices some precision, whereas the Hybrid Ensemble provides the most balanced performance.

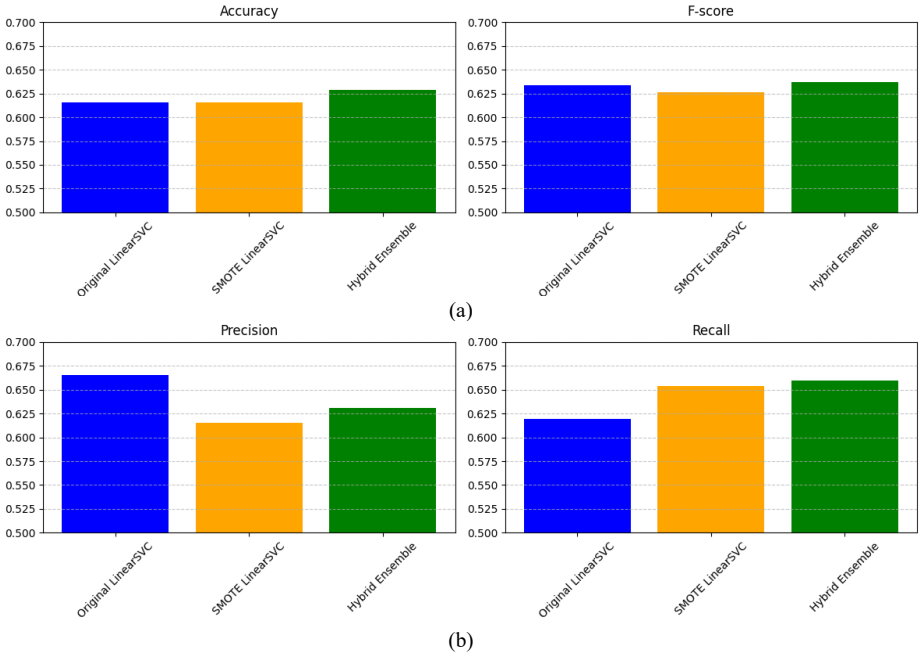

Fig. 8. (A) Class distribution before and after applying SMOTE. (B) Performance comparison of Original LinearSVC, SMOTE LinearSVC, and Hybrid Ensemble models.

B. Confusion Matrix Analysis

The confusion matrices shown in Fig. 9 provide a detailed comparison of the three LinearSVC configurations. The original model demonstrates high precision for majority classes but exhibits considerable misclassification for minority emotions. Several minority classes are frequently predicted as the dominant classes, indicating that the imbalanced training data limits the classifier's ability to learn representative decision boundaries.

After applying SMOTE, the recognition of minority classes improves noticeably. The SMOTE-based model correctly identifies more samples from underrepresented classes; however, the increase in synthetic samples also introduces additional confusion among some majority classes, which leads to a reduction in overall precision.

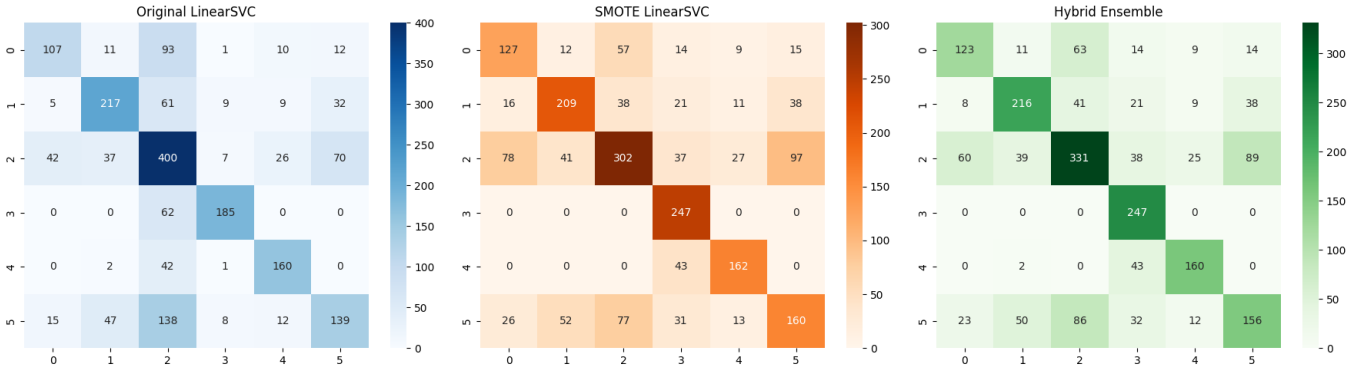

Fig. 9. Confusion matrices of the LinearSVC models: (A) Original model, (B) SMOTE-based model, and (C) Hybrid weighted ensemble.

The Hybrid Ensemble achieves the best balance between these two approaches. The confusion matrix shows a reduction in misclassification errors while maintaining high numbers of correctly classified samples across most emotion categories. This demonstrates that combining the predictions of the original and SMOTE-based models provides a more robust solution to the class imbalance problem than either model individually.

C. Recall Analysis

Fig. 10 compares the recall values obtained by the Original LinearSVC model, the SMOTE-based LinearSVC model, and the Hybrid Weighted Ensemble for each emotion class. The results show that both SMOTE and the Hybrid Ensemble improve recall for minority classes compared with the baseline model.

The Hybrid Ensemble provides the most consistent recall across all emotion classes and achieves the best overall balance between minority and majority class recognition. The most notable result is the perfect recall achieved for Class 3 (Fear), indicating that the proposed weighting strategy effectively enhances the recognition of difficult and underrepresented emotions.

Overall, the additional experiments using LinearSVC confirm the effectiveness of the Hybrid SMOTE-Ensemble concept. Although the XGBoost framework remains the primary contribution of this work, the LinearSVC results demonstrate that the proposed strategy can also improve the performance of other classifiers when dealing with imbalanced Speech Emotion Recognition datasets.

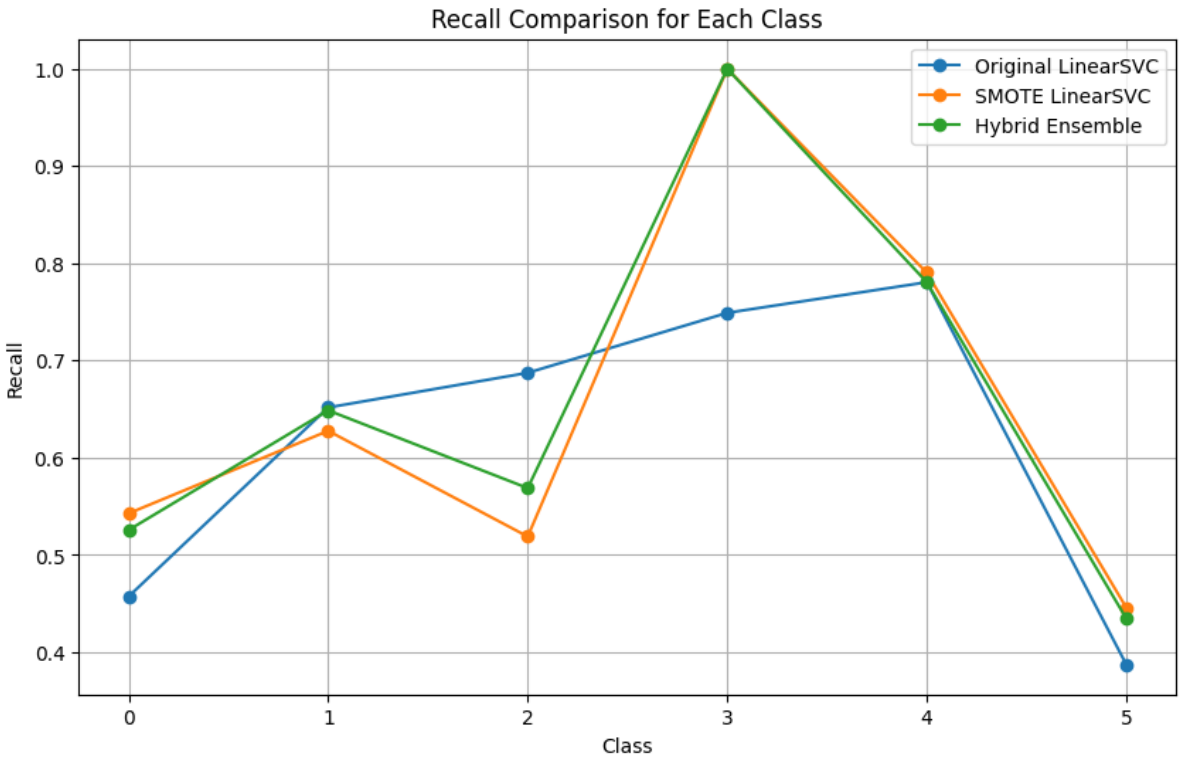

Fig. 10. Recall comparison of the Original LinearSVC model, SMOTE-based LinearSVC model, and Hybrid Weighted Ensemble for each emotion class.

VI. Experimental Results and Discussion

This section presents the experimental results obtained from the reproduction of the original study and the proposed enhancement methods. First, the reproduced results are compared with those reported by Sahu et al. [12] to evaluate the reproducibility of the original framework. Subsequently, the performance of the proposed Hybrid SMOTE-Ensemble Framework is analyzed and compared with the baseline models using different evaluation metrics and confusion matrices.

A. Reproduction Results

The objective of the reproduction experiments was to investigate whether the performance reported by Sahu et al. [12] could be achieved using the publicly available information and implementation details described in their study. The reproduced models were evaluated under three modalities: Audio-only, Text-only, and Audio+Text. Their performance was assessed using Accuracy, F1-score, Precision, and Recall.

Fig. 11 presents the confusion matrix of the Audio-only model reported in [12]. The model classifies emotions into six categories: anger (ang), happiness (hap), sadness (sad), fear (fea), surprise (sur), and neutral (neu). The results show that the model performs well in recognizing sadness and fear, whereas more confusion is observed among anger, happiness, and neutral emotions. This behaviour reflects the inherent difficulty of distinguishing emotions with similar acoustic characteristics.

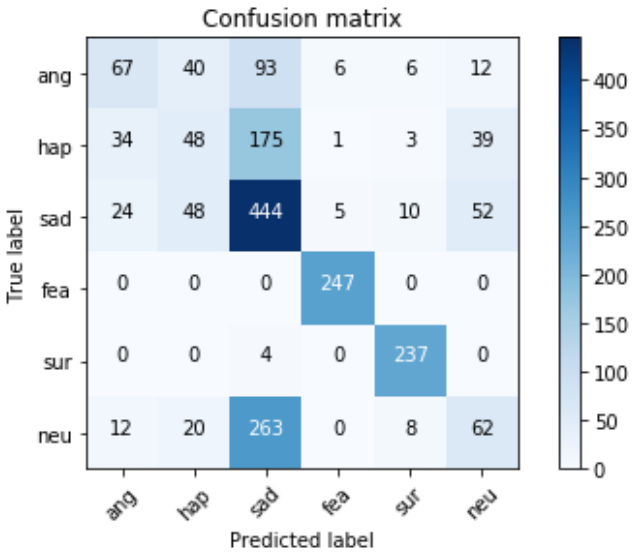

Fig. 11. Results reported by [12] for the audio-only model.

Fig. 12 illustrates the confusion matrix of the Text-only model. The model demonstrates strong performance for sadness and fear, while several misclassifications occur among happy, sad, and neutral emotions. In particular, neutral emotions are frequently predicted as sadness, indicating that textual information alone may not always provide sufficient discriminative features for emotion recognition.

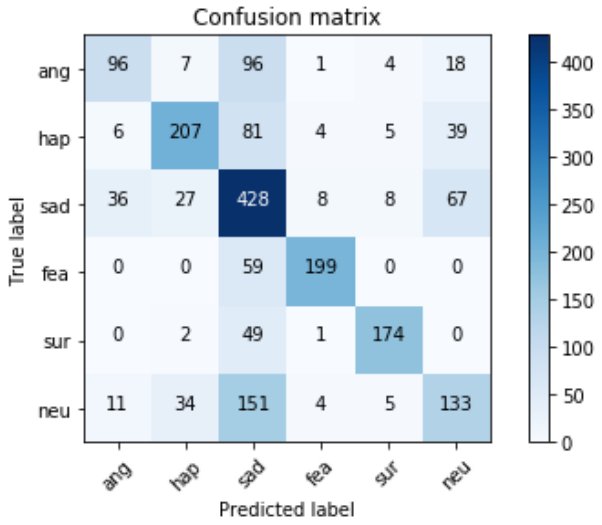

Fig. 12. Results reported by [12] for the text-only model.

Fig. 13 presents the confusion matrix of the Audio+Text model. Compared with the unimodal approaches, the multimodal model provides better recognition performance for most emotion categories by exploiting complementary information from both audio and textual modalities. The model achieves the best performance for sadness and surprise, although some confusion remains between anger and sadness as well as between happy and neutral emotions.

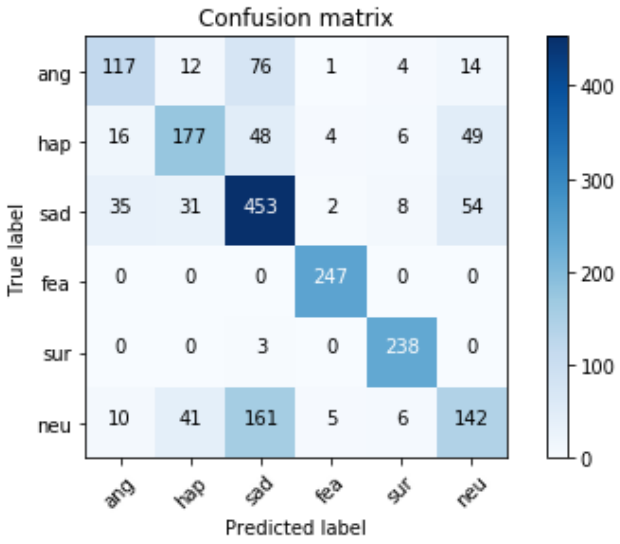

Fig. 13. Results reported by [12] for the audio+text model.

Overall, the reproduced results are in close agreement with those reported by Sahu et al. [12], with most differences remaining within ±1%. Although slight variations are observed for some models and modalities, the general performance trends remain consistent. These differences may be attributed to ambiguities in feature extraction settings, differences in library versions, and the stochastic nature of model training and data partitioning.

B. Results from Enhancements Using Hybrid XGBoost Ensembles

The proposed Hybrid SMOTE-Ensemble Framework was evaluated using the Audio+Text modality. The framework consists of two XGBoost classifiers trained independently on the original imbalanced dataset and the SMOTE-balanced dataset. Their predictions were combined using three ensemble strategies: Simple Averaging, Confidence-Based Selection, and Adaptive Weighting. The performance of these models is summarized in Table I.

TABLE I. PERFORMANCE OF HYBRID XGBOOST ENSEMBLE ENHANCEMENT (AUDIO+TEXT MODALITY)

Model/Ensemble Method	Accuracy (%)	F1-score (%)	Precision (%)	Recall (%)
XGB Original (Baseline)	54.9	54.9	71.9	50.7
XGB SMOTE	56.4	59.4	62.1	58.7
Simple Average Ensemble	57.7	59.3	70.5	55.3
Confidence-Based Ensemble	57.0	58.5	70.9	54.4
Adaptive Weighted Ensemble	57.4	58.9	71.1	54.9

The results demonstrate that training XGBoost on SMOTE-balanced data improves the classification performance compared with the original imbalanced training set. The XGB SMOTE model achieved an accuracy of 56.4% and an F1-score of 59.4%, compared with 54.9% accuracy and 54.9% F1-score obtained by the original XGBoost model. This improvement indicates that balancing the training data enables the classifier to better recognize underrepresented emotion classes.

The ensemble methods further improve the overall performance. The Simple Average Ensemble achieved the highest accuracy of 57.7%, while the Adaptive Weighted Ensemble achieved competitive results with an accuracy of 57.4% and an F1-score of 58.9%. The Confidence-Based Ensemble also outperformed the baseline model and achieved balanced performance across the evaluation metrics.

Overall, the results confirm that the proposed Hybrid SMOTE-Ensemble Framework improves emotion recognition performance on imbalanced data. The combination of SMOTE-based data balancing and ensemble learning provides a more robust and balanced classifier compared with a single model trained on the original dataset.

C. Hybrid LinearSVC Ensemble Enhancement Results

To further investigate the effectiveness of the proposed enhancement strategy, additional experiments were conducted using a Linear Support Vector Classifier (LinearSVC). Three different configurations were evaluated: the Original LinearSVC trained on the imbalanced dataset, the SMOTE-based LinearSVC trained on the balanced dataset, and the Hybrid Weighted Ensemble combining the predictions of both models.

The results presented in Tables II and III demonstrate that applying SMOTE improves the recall of minority emotion classes, although this improvement is accompanied by a slight reduction in precision for some majority classes. The Original LinearSVC model achieved the highest precision; however, it suffered from lower recall, particularly for underrepresented emotions.

TABLE II. PERFORMANCE METRICS OF TECHNIQUES USED WITH LINEARSVC

Metric	Original LinearSVC	SMOTE LinearSVC	Hybrid Ensemble
Accuracy	0.616	0.616	0.629
F-score	0.634	0.626	0.637
Precision	0.665	0.615	0.631
Recall	0.619	0.654	0.660

TABLE III. RECALL VALUES FOR EACH CLASS OF TECHNIQUES USED WITH LINEARSVC

Class	Original LinearSVC	SMOTE LinearSVC	Hybrid Ensemble
0	0.457	0.543	0.526
1	0.652	0.628	0.649
2	0.687	0.519	0.569
3	0.749	1.000	1.000
4	0.780	0.790	0.780
5	0.387	0.446	0.435

The Hybrid Weighted Ensemble achieved the best overall performance among the evaluated LinearSVC models. The ensemble obtained the highest accuracy and F1-score while maintaining balanced precision and recall values. These results indicate that combining the original and SMOTE-based classifiers enables the model to exploit the strengths of both

approaches and provides a more robust solution for imbalanced Speech Emotion Recognition.

Furthermore, the recall analysis confirms that the Hybrid Ensemble consistently improves the recognition of minority emotion classes. The results suggest that the proposed Hybrid SMOTE-Ensemble strategy is not limited to XGBoost and can also enhance the performance of other machine learning classifiers such as LinearSVC.

Overall, the additional LinearSVC experiments support the effectiveness and generalization capability of the proposed enhancement framework for imbalanced Speech Emotion Recognition tasks.

D. Analysis of Results

The reproduction experiments demonstrated a high degree of agreement with the results reported by Sahu et al. [12]. Most evaluation metrics differed by less than 1%, indicating that the proposed reproduction methodology successfully replicated the main findings of the original study. The small differences observed for some models may be attributed to ambiguities in feature extraction settings, differences in library implementations, and the inherent stochastic nature of machine learning and deep learning training processes.

The results obtained from the proposed Hybrid SMOTE-Ensemble Framework further demonstrate the effectiveness of combining data balancing and ensemble learning for Speech Emotion Recognition on imbalanced datasets. Training XGBoost on SMOTE-balanced data improved both Accuracy and F1-score compared with the original imbalanced model. In particular, the F1-score increased from 54.9% to 59.4%, indicating that balancing the training data enables the classifier to better recognize underrepresented emotion classes.

The ensemble methods provided additional improvements over the baseline XGBoost model. Among the evaluated approaches, the Simple Average Ensemble achieved the highest accuracy of 57.7%, while the Adaptive Weighted Ensemble achieved competitive results with balanced Precision, Recall, and F1-score values. These findings suggest that combining classifiers trained on original and balanced datasets produces a more robust model than relying on a single classifier.

The additional experiments using LinearSVC further support this observation. Although the overall improvement was less pronounced than that obtained with XGBoost, the Hybrid Weighted Ensemble consistently outperformed the baseline LinearSVC model and improved the recognition of minority emotion classes. These results indicate that the proposed Hybrid SMOTE-Ensemble strategy is not limited to a specific classifier and can be generalized to other machine learning models.

E. Robustness and Limitations

The robustness of the experiments was improved by fixing the random seed (`random_state = 42`) during train-test splitting, SMOTE oversampling, and model training whenever possible. This ensures that the experiments can be reproduced and that the obtained results are not caused by random variations in data partitioning or model initialization.

The reproduced results remained close to those reported by Sahu et al. [12], demonstrating that the original findings are generally robust to minor differences in implementation details and software environments. In addition, the proposed Hybrid SMOTE-Ensemble Framework consistently improved performance across different evaluation metrics, indicating that the enhancement strategy is stable and effective for imbalanced Speech Emotion Recognition tasks.

Despite these encouraging results, several limitations should be acknowledged. First, the enhancement strategy was evaluated only using XGBoost and LinearSVC classifiers. The impact of the proposed framework on other classifiers, such as Random Forest, SVM, MLP, and LSTM, remains an open research direction. Second, some assumptions were required during the reproduction process because the original study did not explicitly specify all implementation details, particularly those related to feature extraction and label preprocessing. Consequently, the reproduced experiments may not represent an exact replication of the original experimental setup.

F. Ethical Considerations and Broader Impact

Speech Emotion Recognition systems have the potential to improve human-computer interaction, customer service, healthcare, and assistive technologies. However, incorrect emotion prediction may negatively affect users, particularly in sensitive applications such as mental health assessment or security systems. Therefore, developing robust and fair emotion recognition systems is an important ethical consideration.

Class imbalance is one of the factors that may introduce bias into Speech Emotion Recognition models because minority emotions are often underrepresented during training. The proposed Hybrid SMOTE-Ensemble Framework attempts to mitigate this issue by improving the recognition of minority emotion classes through balanced data generation and ensemble learning. Nevertheless, synthetic data generated by oversampling techniques should be used carefully to avoid introducing artificial patterns or hidden biases into the training process.

Finally, this study highlights the importance of reproducibility and transparent reporting in machine learning research. Providing clear implementation details and thoroughly evaluating enhancement strategies contribute to the development of more reliable, fair, and trustworthy Speech Emotion Recognition systems.

VII. Conclusion

This study had two main objectives. The first was to reproduce the multimodal Speech Emotion Recognition framework proposed by Sahu et al. [12] and evaluate the reproducibility of their reported results. The second was to enhance the original framework by introducing a Hybrid SMOTE-Ensemble strategy based on XGBoost and LinearSVC classifiers to address the class imbalance problem in the IEMOCAP dataset.

The reproduction experiments demonstrated a high level of agreement with the original study, with most evaluation metrics differing by less than 1%.

These results confirm the validity of the original findings and highlight the importance of transparent reporting and reproducible research practices in Speech Emotion Recognition.

The enhancement experiments showed that balancing the training data using SMOTE improves the performance of emotion classification models, particularly in terms of F1-score and Recall. Furthermore, combining models trained on the original and SMOTE-balanced datasets through ensemble learning provided additional improvements in overall performance. The proposed Hybrid SMOTE-Ensemble Framework achieved more balanced and robust performance than the baseline models, demonstrating the effectiveness of integrating data balancing with ensemble learning for imbalanced Speech Emotion Recognition tasks.

The additional experiments using LinearSVC further confirmed the generalization capability of the proposed framework. Although the performance improvements varied across classifiers, the Hybrid Ensemble consistently provided a better balance between Precision, Recall, and F1-score than individual models.

Overall, this study demonstrates that advanced oversampling techniques and ensemble learning strategies can effectively mitigate class imbalance and improve Speech Emotion Recognition performance. The findings also emphasize the importance of reproducibility and systematic experimentation in developing reliable and robust emotion recognition systems.

VIII. Future Research and Implementations

Several directions can be explored to further improve the proposed Hybrid SMOTE-Ensemble Framework:

- Investigate the effectiveness of the proposed framework using additional classifiers such as Random Forest, Support Vector Machine, Multi-Layer Perceptron, and Long Short-Term Memory networks.
- Evaluate the Hybrid SMOTE-Ensemble strategy using different modalities, including Audio-only and Text-only settings, as well as more advanced multimodal fusion approaches.
- Explore more sophisticated oversampling techniques, such as SISMOTE and FC-SMOTE, or investigate generative approaches based on Generative Adversarial Networks (GANs) for data augmentation.
- Perform more extensive hyperparameter optimization for both the base classifiers and the ensemble strategies to further improve classification performance.
- Investigate the proposed framework on other Speech Emotion Recognition datasets with different class distributions to evaluate its robustness and generalization capability.

These directions may contribute to the development of more accurate, robust, and fair Speech Emotion Recognition systems capable of handling the challenges associated with imbalanced emotional datasets.

References

[1] Abdelhamid, A. A., El-Kenawy, E. S. M., Alotaibi, B., Amer, G. M., Abdelkader, M. Y., Ibrahim, A., & Eid, M. M. (2022). Robust speech emotion recognition using CNN+ LSTM based on stochastic fractal search optimization algorithm. IEEE Access, 10, 49265–49284.

[2] Akbani, R., Kwek, S., & Japkowicz, N. (2004). Applying support vector machines to imbalanced datasets. In J.-F. Boulicaut, F. Esposito, F. Giannotti, & D. Pedreschi (Eds.), Lecture Notes in Computer Science: Vol. 3201. Machine Learning: ECML 2004 (pp. 39–50). Springer-Verlag Berlin Heidelberg.

[3] Atmaja, B. T., & Sasou, A. (2022). Effects of data augmentations on speech emotion recognition. Sensors, 22(16), 5941.

[4] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4), 335–359. https://doi.org/10.1007/s10579-008-9076-6

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

[6] Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEBoost: Improving prediction of the minority class in boosting. In Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2003) (pp. 107–119).

[7] Fan, Z., Wu, Y., Zhou, C., Zhang, X., & Tao, Z. (2021). Class-imbalanced voice pathology detection and classification using fuzzy cluster oversampling method. Applied Sciences, 11(8), 3450.

[8] Fitzgerald, D. (2010). Harmonic/percussive separation using median filtering. In Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10).

[9] Liang, X., Jiang, H., Xu, W., & Zhou, Y. (2023). Gaussian-smoothed imbalance data improves speech emotion recognition. arXiv preprint arXiv:2302.08650.

[10] Liu, Z.-T., Wu, B.-H., Li, D.-Y., Xiao, P., & Mao, J.-W. (2020). Speech emotion recognition based on selective interpolation synthetic minority over-sampling technique in small sample environment. Sensors, 20(8), 2297.

[11] Meng, T., Shou, Y., Ai, W., Yin, N., & Li, K. (2023). Deep imbalanced learning for multimodal emotion recognition in conversations. arXiv preprint arXiv:2312.06337.

[12] Sahu, G. (2019). Multimodal speech emotion recognition and ambiguity resolution. arXiv preprint arXiv:1904.06022v1 [cs.LG].

[13] Sondhi, M. M. (1968). New methods of pitch extraction. IEEE Transactions on Audio and Electroacoustics, 16(2), 262–266.

[14] Weiss, G. M. (2004). Mining with rarity: A unifying framework. ACM SIGKDD Explorations Newsletter, 6(1), 7–19.

[15] Zhao, J., Dong, W., Shi, L., Qiang, W., Kuang, Z., Xu, D., & An, T. (2022). Multimodal feature fusion method for unbalanced sample data in social network public opinion. Sensors, 22(15), 5528.

[16] Zhou, Z.-H., & Liu, X.-Y. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering, 18(1), 63–77.

[17] Mashagba, H. A., Rahim, H. A., Mohd Yasin, M. N., Jamaluddin, M. H., Islam, M. T., Abdulkawi, W. M., ... & Al-Bawri, S. S. (2024). Four-port dual-band textile MIMO antenna for biomedical health monitoring systems: on-body mutual coupling reduction characterization. *Physica Scripta*, *99*(9), 095026.

[18] Owida, H. A., El-Fattah, I. A., Abuowaida, S., Alshdaifat, N., Mashagba, H. A., Abd Aziz, A. B., ... & Al-Bawri, S. S. (2025). A deep learning-based dual-branch framework for automated skin lesion segmentation and classification via dermoscopic Images. Scientific Reports, 15(1), 37823.

[19] Rahim, H. A., Mashagba, H. A., Yahaya, N. Z., Yasin, M. N. M., Jamaluddin, M. H., Jusoh, M., ... & Abdulmalek, M. (2021, December). 5G Millimeter Wave Wearable Antenna: State-Of-the-Art and Current Challenges. In *2021 IEEE Asia-Pacific Conference on Applied Electromagnetics (APACE)* (pp. 1-4). IEEE.

[20] A. Al-Mashagba, L., Sumari, P., Mashagba, M., Al Abri, F., Mashagba, H. A., Abd Aziz, A. B., & Al-Bawri, S. S. (2026). Machine learning and Geographic Information Systems (GIS)–based model for road crash severity prediction and behavioural pattern analysis. Traffic Injury Prevention, 1-10.

[21] Mashagba, H. A., Rahim, H. A., Yasin, M. N. M., Jamaluddin, M. H., Narongkul, S., Ramli, N. H., & Zahid, L. (2024, November). Two-Port Hexagonal Bar Slotted Dual-Band MIMO Textile Antenna for WBAN and 5G Applications. In 2024 International Symposium on Antennas and Propagation (ISAP) (pp. 1-2). IEEE.

[22] Rawash, Y. Z., Al-Naami, B., Alfraihat, A., & Owida, H. A. (2024). Advanced Low-Pass Filters for Signal Processing: A Comparative Study on Gaussian, Mittag-Leffler, and Savitzky-Golay Filters. Mathematical Modelling of Engineering Problems, 11(7).

[23] Al-Naami, B., Fraihat, H., Owida, H. A., Al-Hamad, K., De Fazio, R., & Visconti, P. (2022). Automated detection of left bundle branch block from ECG signal utilizing the maximal overlap discrete wavelet transform with ANFIS. Computers, 11(6), 93.

[24] Jawwad, A. K. A., Turab, N., Al-Mahadin, G., Owida, H. A., & Al-Nabulsi, J. (2024). A perspective on smart universities as being downsized smart cities: a technological view of internet of thing and big data. Indonesian Journal of Electrical Engineering and Computer Science, 35(2), 1162-1170.

[25] Abuowaida, S., Owida, H., Alsekait, D., Alshdaifat, N., Abdelminaam, D., & Alshinwan, M. (2025). UltraSegNet: A hybrid deep learning framework for enhanced breast cancer segmentation and classification on ultrasound images. *Computers, Materials, & Continua*, *83*(2), 3303.

[26] Owida, H. A., Hemied, O. S. M., Alkhawaldeh, R. S., Alshdaifat, N. F. F., & Abuowaida, S. F. A. (2022). Improved deep learning approaches for covid-19 recognition in ct images. *Journal of Theoretical and Applied Information Technology*, *100*(13), 4925-4931.

[27] Abuowaida, S. F. A., Chan, H. Y., Alshdaifat, N. F. F., & Abualigah, L. (2021). A novel instance segmentation algorithm based on improved deep learning algorithm for multi-object images. *Jordanian Journal of Computers and Information Technology (JJCIT)*, *7*(01).

[28] Abuowaida, S. U. H. A. I. L. A., Elsoud, E. S. R. A. A., Al-Momani, A. D. A. I., Arabiat, M. O. H. A. M. M. A. D., Owida, H. A., ALSHDAIFAT, N., & CHAN, H. Y. (2023). Proposed enhanced feature extraction for multi-food detection method. Journal of Theoretical and Applied Information Technology, 101(24), 8140-8146.

[29] Turki, H. M., Al Daoud, E., Samara, G., Alazaidah, R., Qasem, M. H., Aljaidi, M., ... & Alshdaifat, N. (2025). Arabic fake news detection using hybrid contextual features. International Journal of Electrical & Computer Engineering (2088-8708), 15(1).

[30] Arabiat, M. O. H. A. M. M. A. D., Abuowaida, S. U. H. A. I. L. A., Al-Momani, A. D. A. I., Alshdaifat, N. A. W. A. F., & Chan, H. Y. (2024). Depth estimation method based on residual networks and se-net model. Journal of Theoretical and Applied Information Technology, 102(3), 866-871.

[31] Alidmat, O., Owida, H. A., Yusof, U. K., Almaghthawi, A., Altalidi, A., Alkhawaldeh, R. S., ... & Alshaqsi, J. (2025). Simulation of Crowd Evacuation in Asymmetrical Exit Layout Based on Improved Dynamic Parameters Model. *IEEE Access*, *13*, 7512-7525.

[32] Ehsan, A., Iqbal, Z., Abuowaida, S., Aljaidi, M., Zia, H. U., Alshdaifat, N., & Alshammry, N. K. (2024). Enhanced anomaly detection in ethereum: Unveiling and classifying threats with machine learning. IEEE Access, 12, 176440-176456

[33] Iqtait, M., Ababneh, J., Rasmi, M., Abu-Jassar, A., & Abuowaida, S. (2024). Enhancing facial feature detection: hybrid active shape and active appearance model (HASAAM). IEEE Access, 12, 189590-189610.

[34] Owida, H. A., Al-Nabulsi, J., Al Hawamdeh, N., Abuowaida, S., Salah, Z., & Elsoud, E. A. (2024, November). Deep learning-based kidney disease classification using ResNet50: Enhancing diagnostic accuracy. In 2024 Second Jordanian International Biomedical Engineering Conference (JIBEC) (pp. 126-129). IEEE.

[35] Baniata, M., Abuowaida, S., Aljaidi, M., Kharabsheh, M., Alsarhan, A., & Alsuwaylimi, A. A. (2025). A Multi-Modal Attention-Guided Network for Alzheimer's Disease Classification Using Deep Learning. *Engineering, Technology & Applied Science Research*, *15*(5), 27150-27158.

[36] Aburomman, A., Abuowaida, S., Al-Momani, A., Al Henawi, E., Elsoud, E. A., Salah, Z., ... & Khouj, M. (2024, December). Automated detection of Alzheimer's disease using EfficientNet-B3 architecture with transfer learning. In 2024 25th International Arab Conference on Information Technology (ACIT) (pp. 1-6). IEEE.

[37] Malkawi, M. M., Khaleel, A., Bataineh, M., Alkhatib, R., Nayfeh, O., Alqaraleh, M., & Abuowaida, S. (2025, April). Enhancing Autism Spectrum Disorder Detection: A Novel Ensemble Learning Framework with Multi-Model Integration. In 2025 1st International Conference on Computational Intelligence Approaches and Applications (ICCIAA) (pp. 1-6). IEEE.

[38] Abuowaida, S., Alnsour, Y., Salah, Z., Alazaidah, R., Al-Batah, M. S., Alzboon, M. S., ... & Moh'd, B. A. H. (2025). Hybrid Ensemble Architecture for Brain Tumor Segmentation Using EfficientNetB4-MobileNetV3 with Multi-Path Decoders. DATA & METADATA Учредители: Salud, Ciencia y Tecnologia, 4, 374.

[39] Salah, Z., Abu Owida, H., Abu Elsoud, E., Alhenawi, E., Abuowaida, S., & Alshdaifat, N. (2024). An effective ensemble approach for preventing and detecting phishing attacks in textual form. Future Internet, 16(11), 414.

[40] Owida, H. A., Hassan, M. R., Ali, A. M., Alnaimat, F., Al Sharah, A., Abuowaida, S., & Alshdaifat, N. (2024). The performance of artificial intelligence in prostate magnetic resonance imaging screening. *International Journal of Electrical & Computer Engineering (2088-8708)*, *14*(2).