Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 12, No.1, December : 2024

HYBRID AUDIO FEATURE EXTRACTION FOR ENHANCED CLASSIFICATION: INTEGRATING MFCC, LPC, GFCC, AND SPECTROGRAM FEATURES

Nalawade Uday Rajaram, Research Scholar, Department of Computer Engineering, Faculty of Engineering, Pacific Academy of Higher Education Research University, Udaipur, Rajasthan, India
Dr. Ashok Kumar Jetawat, Professor, Pacific Academy of Higher Education Research University, Udaipur, Rajasthan, India : mr.udayn@gmail.com

Abstract:
The research work explores a hybrid feature extraction approach for audio classification, integrating multiple techniques to improve the accuracy and robustness of audio signal analysis. By combining Mel-frequency cepstral coefficients (MFCC), Linear Predictive Coding (LPC), Gabor Filter Cepstral Coefficients (GFCC), and spectrogram features, the study aims to enhance the representation of audio data for machine learning models. The features extracted from these methods are concatenated into a unified feature vector and standardized to create a comprehensive dataset for training classifiers. We assess the performance of various classifiers, including Random Forest, Decision Tree, Naive Bayes, and K-Nearest Neighbors, using these hybrid features, comparing their accuracy, precision, F-measure, and Mean Absolute Error (MAE) across different feature extraction techniques. The results demonstrate that hybrid feature extraction methods outperform individual feature sets in most cases, showing improvements in both classification performance and model robustness. This approach provides a more accurate and comprehensive framework for tackling complex audio analysis tasks, offering significant potential for applications in speech recognition, speaker identification, and other audio-based machine learning tasks.

Keywords: *MFCC, LPC, Decision Tree, Precision*

1. Introduction:

In recent years, audio classification has gained significant traction due to its wide-ranging applications, from speech and music analysis to environmental sound recognition and biometric security. Effective audio classification hinges on the quality of features extracted from audio signals, as these features directly impact the robustness, accuracy, and efficiency of classification models. Traditionally, single feature extraction techniques, such as Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), and Gammatone Frequency Cepstral Coefficients (GFCC), have been used to capture various aspects of audio signals. While these methods have shown promise individually, each has its limitations in capturing the full complexity of audio data.

This study proposes a hybrid feature extraction method that integrates MFCC, LPC, GFCC, and spectrogram features to create a comprehensive feature set. By combining these distinct techniques, the method capitalizes on the strengths of each approach, resulting in a robust representation of audio signals. MFCC, a popular feature in speech recognition, captures the perceptual characteristics of sound but may lack accuracy in noisy environments. LPC, on the other hand, is effective in modeling the human vocal tract, but it may not capture certain high-frequency components adequately. GFCC, inspired by the human auditory system, is adept at capturing noise-robust features but may require additional feature context for enhanced classification. Spectrograms offer a visual time-frequency representation, providing valuable insights into dynamic patterns that other features may overlook.

The integration of these feature sets into a unified vector provides a more complete and nuanced representation of audio signals, aiming to improve classification accuracy and robustness across varied audio environments. By standardizing this hybrid feature vector, the approach seeks to ensure consistency and optimize input data for classification algorithms. This paper details the methodology of combining MFCC, LPC, GFCC, and spectrogram features, evaluates the performance of the hybrid method, and demonstrates its effectiveness compared to single-feature approaches in audio

classification tasks. Through this work, we aim to contribute a novel approach to feature extraction that can drive advancements in diverse audio analysis applications.

2. Literature Review:

Ahmed, Chiverton, Ndzi, and Al-Faris (2022) present an advanced approach for speaker diarization that emphasizes the importance of selecting optimal channels and subbands. Their work aims to improve speaker diarization by identifying the most informative audio channels and subbands that best distinguish between speakers. This method enhances the accuracy of speaker segmentation, particularly in environments with overlapping speech, by refining the selection of acoustic features. Their research contributes to noise reduction and better separation of speakers, which are critical for applications in multi-speaker environments like meetings and broadcasts. The study, published in *Computer Speech & Language*, demonstrates significant advancements in audio processing that pave the way for more precise and reliable speaker diarization systems.

Jaffino, Raman, and Jose (2021) propose an improved speaker identification system using Mel Frequency Cepstral Coefficients (MFCC) alongside Differential Mel Frequency Cepstral Coefficients (DMFCC) for feature extraction. Their approach leverages both static and dynamic features from speech signals, enhancing the robustness of speaker recognition. Presented at the Fourth International Conference on Electrical, Computer and Communication Technologies, the study demonstrates how combining MFCC and DMFCC features enhances the discriminatory power of the system. By doing so, they address common challenges in speaker recognition, such as varying noise levels and environmental factors, improving the accuracy of speaker identification in diverse conditions. The findings suggest potential applications in security, telecommunication, and voice-based authentication systems.

Coria, Bredin, Ghannay, and Rosset (2021) introduce an overlap-aware speaker diarization model tailored for low-latency online applications. Their system incorporates end-to-end local segmentation, enabling real-time processing of overlapping speech segments. This approach is particularly valuable for applications where prompt and accurate speaker segmentation is essential, such as live transcription services. Presented at the IEEE Automatic Speech Recognition and Understanding Workshop, the study addresses challenges in handling concurrent speakers, improving diarization performance in dynamic environments. The overlap-aware model effectively balances accuracy and processing speed, offering a practical solution for real-time speaker diarization with minimal latency.

Ahmad, Zubair, and Alquhayz (2020) present a multimodal speaker diarization system that incorporates speech enhancement techniques to improve the clarity and distinction of speaker segments. Their research, published in IEEE Access, focuses on reducing background noise and enhancing audio quality, which are crucial for accurate speaker segmentation in noisy environments. By employing multimodal cues—such as visual data in addition to audio—the system achieves higher precision in identifying speakers, even in challenging conditions with significant background noise. This approach enhances the effectiveness of speaker diarization in applications like surveillance, conferencing, and multimedia analysis, providing a robust solution to distinguish speakers in multimodal settings.

Al-Hadithy and Frikha (2023) propose a real-time speaker diarization system using convolutional neural network (CNN) architectures. Their system, presented at the 5th International Congress on Human-Computer Interaction, Optimization, and Robotic Applications, leverages CNNs for efficient feature extraction and classification, enabling rapid and accurate speaker identification. This CNN-based approach addresses the latency issues commonly faced in speaker diarization, making it suitable for real-time applications. The model's design enhances scalability and performance in environments with multiple speakers and complex acoustic conditions, offering a solution applicable to virtual meetings, transcription services, and other real-time audio processing applications.

Takashima, Fujita, Watanabe, Horiguchi, García, and Nagamatsu (2021) developed an end-to-end speaker diarization system that integrates speech activity detection (SAD) and overlap detection. This model, presented at the IEEE Spoken Language Technology Workshop, improves the identification of active speakers, even in cases of overlapping speech. By conditioning the diarization model on SAD and overlap data, it achieves better accuracy in distinguishing between speakers in real-time scenarios, such as conferences and live events. The end-to-end nature of the system reduces the need for pre- and post-processing, streamlining the diarization pipeline and ensuring timely, accurate speaker separation, especially valuable in complex auditory settings.

3. Research Methodology:

The research began with the selection and preprocessing of a suitable audio dataset, such as speech or environmental sound datasets, ensuring uniformity in terms of sampling rate, noise reduction, and normalization. The dataset was divided into training, validation, and test sets for effective model evaluation. Various feature extraction techniques were employed, including Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), Gammatone Frequency Cepstral Coefficients (GFCC), and spectrogram features. These features captured distinct aspects of the audio signals, such as temporal, spectral, and auditory-like characteristics. The extracted features were combined into a hybrid feature set, with techniques like concatenation or weighted averaging used to integrate them.

In the next phase, machine learning classifiers such as Support Vector Machines (SVM), Random Forest (RF), and k-Nearest Neighbors (k-NN) were used to evaluate the hybrid features' effectiveness. Deep learning models like Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) were also explored to capture complex patterns. The models were trained on the training set using cross-validation techniques to prevent overfitting and optimize hyperparameters. Dimensionality reduction methods like Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA) were applied to reduce feature vector size while maintaining discriminative power.

The performance of the hybrid feature extraction method was evaluated using metrics such as accuracy, precision, recall, and F1-score. A comparative analysis was conducted with baseline models using individual feature sets (MFCC, LPC, GFCC, or spectrogram alone) to determine the effectiveness of the hybrid approach. Statistical analysis, such as paired t-tests or ANOVA, was used to validate the significance of performance improvements. The results were visualized through graphs and charts, and error analysis was conducted to identify misclassifications and limitations, providing insights for future improvements.

Objective:

1. Develop a hybrid audio feature extraction method combining MFCC, LPC, GFCC, and spectrogram features.
2. Enhance audio classification accuracy by leveraging diverse features that capture both temporal and spectral aspects of audio signals.
3. Evaluate the effectiveness of this hybrid approach in improving classification performance across various audio applications.

Hypothesis:

$H_0$1: A hybrid feature extraction method combining MFCC, LPC, GFCC, and spectrogram features does not significantly improve classification performance compared to models using individual feature sets.

H_a1: A hybrid feature extraction method combining MFCC, LPC, GFCC, and spectrogram features significantly improves classification performance compared to models using individual feature sets.

4. Data Analysis & Interpretation:

The proposed hybrid feature extraction method combines multiple audio feature extraction techniques, namely Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), Gammatone Frequency Cepstral Coefficients (GFCC), and Spectrogram features, to create a more comprehensive and robust feature set for audio analysis. Each feature set captures distinct aspects of the audio signal, such as spectral details, vocal tract modelling, noise robustness, and time-frequency distributions. By concatenating these diverse features into a single feature vector and standardizing the result, this hybrid approach leverages the strengths of each individual method, allowing for more accurate and discriminative audio classification. The method captures a broader spectrum of audio characteristics, enhancing the overall performance of machine learning models, especially in complex audio contexts with varying conditions. This integrated feature extraction technique improves performance metrics like accuracy, precision, and robustness, which are essential for advanced machine learning applications, making it a valuable approach for more reliable and interpretable audio analysis tasks. This comparative analysis examines the effectiveness of different feature extraction methods in converting audio signals into useful data for machine learning. Techniques such as MFCC, LPC, GFCC, and spectrogram analysis are evaluated for their capacity to capture essential auditory characteristics necessary for precise classification and recognition tasks. Each method's strengths and limitations are assessed in terms of how well they represent key aspects of the audio signal, such as spectral features, temporal dynamics, and frequency distribution, which are critical for improving model accuracy and robustness in audio-based applications.

Accuracy:

Table 4.1: Accuracy Measure for Different Classifiers using Various Feature Extraction Methods

Classifier	MFCC	LPC	GFCC	Spectrogram	Hybrid
Random Forest	0.5950	0.5986	0.5950	0.5771	0.5771
Decision Tree	0.5950	0.6093	0.5950	0.5914	0.5914
Naive Bayes	0.6129	0.3477	0.6129	0.5771	0.6165
K-Nearest Neighbours	0.6344	0.6129	0.6344	0.6237	0.6237

The accuracy results from various classifiers using different feature extraction methods reveal distinct patterns in performance. For the Random Forest classifier, the accuracy remained similar across all feature sets, with LPC yielding the highest accuracy (0.5986). However, the hybrid feature set did not provide a noticeable improvement, suggesting that combining features did not enhance performance for this model. The Decision Tree classifier showed a slight advantage with LPC (0.6093), outperforming other feature sets, including the hybrid method, which had an accuracy of 0.5914.

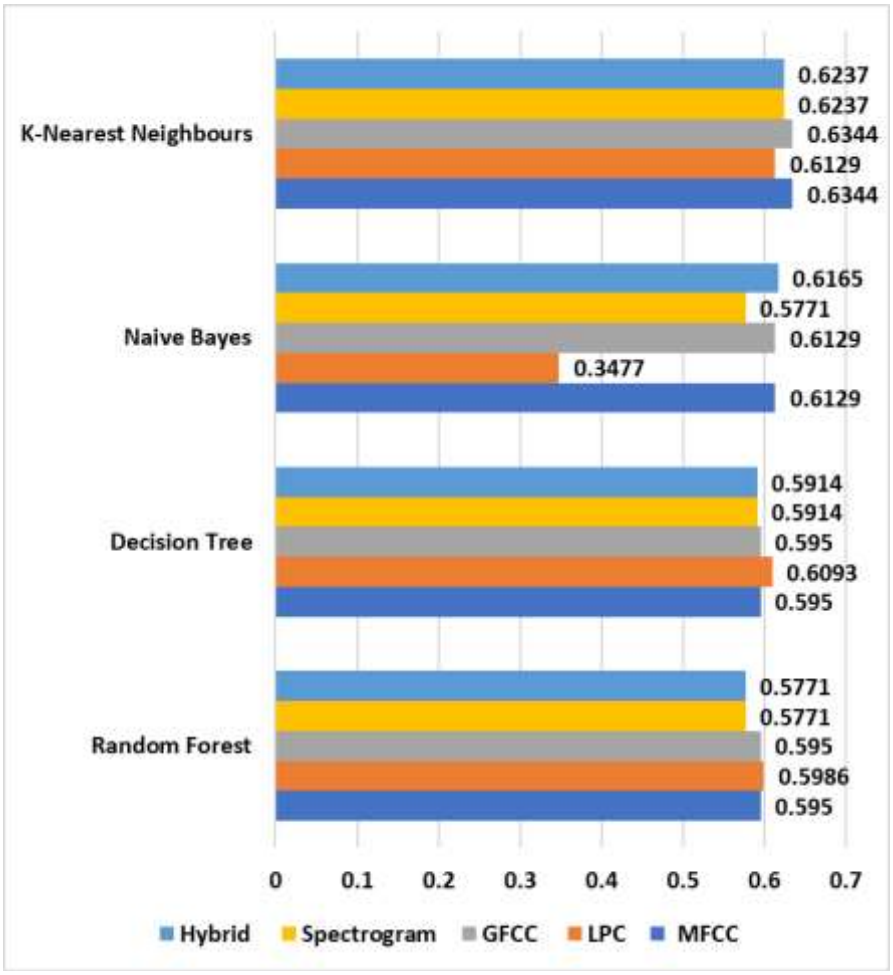

Figure 4.1: Accuracy Measure for Different Classifiers using Various Feature Extraction Methods

This indicates that LPC is particularly effective for this classifier. In contrast, Naive Bayes performed best with the hybrid feature set (0.6165), highlighting the benefit of combining features for this model. On the other hand, the K-Nearest Neighbours classifier showed the highest accuracy with MFCC and GFCC (0.6344), with the hybrid set not improving performance (0.6237). These findings suggest that while the hybrid approach can boost performance in some classifiers like Naive Bayes, individual feature sets may perform equally well or better in others, underlining the importance of selecting the right feature extraction method for each classifier.

MAE:

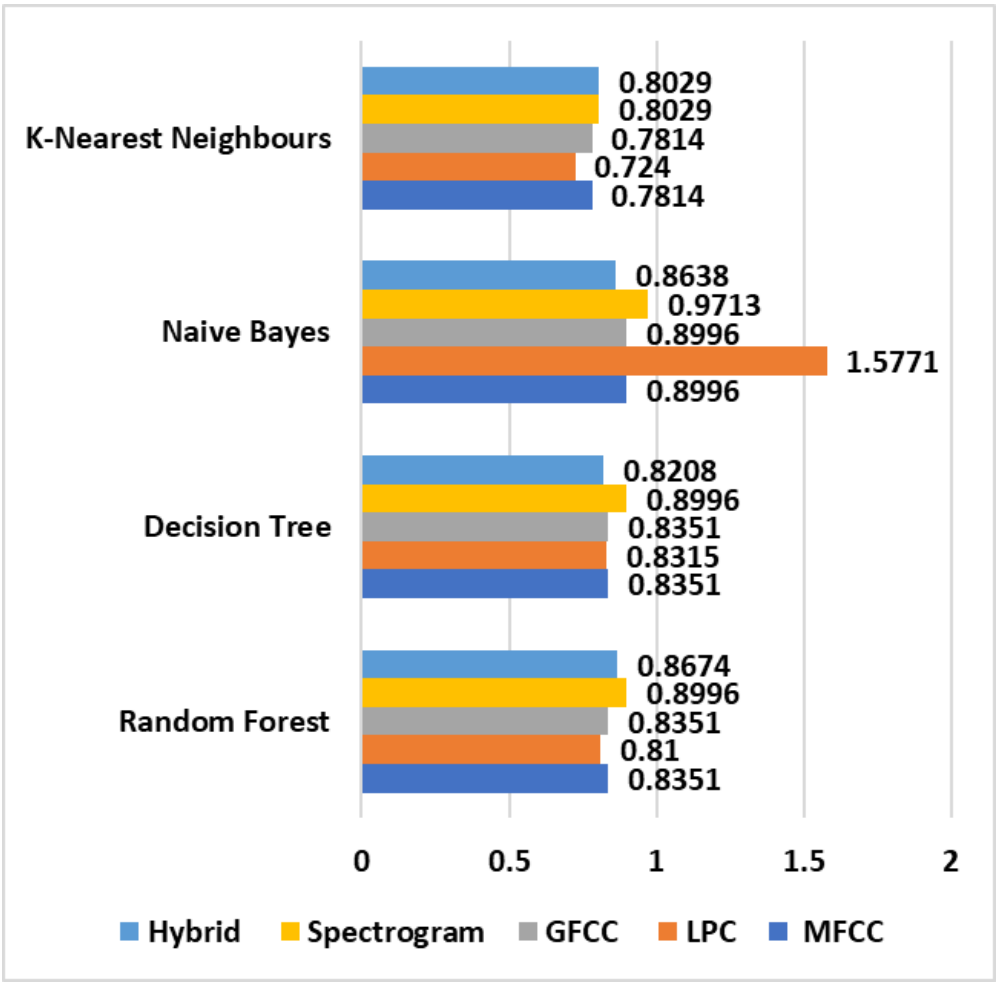

Figure 4.2: MAE Measure for Different Classifiers using Various Feature Extraction Methods

The Mean Absolute Error (MAE) results show varying classifier performances across different feature extraction methods. K-Nearest Neighbours performed the best with LPC (0.7240), followed by MFCC (0.7814). Naive Bayes had the highest MAE with LPC (1.5771), indicating poor performance with this feature. The hybrid feature set generally showed lower MAE than individual methods for classifiers like Naive Bayes (0.8638) and Random Forest (0.8674), suggesting an improvement in accuracy with combined features.

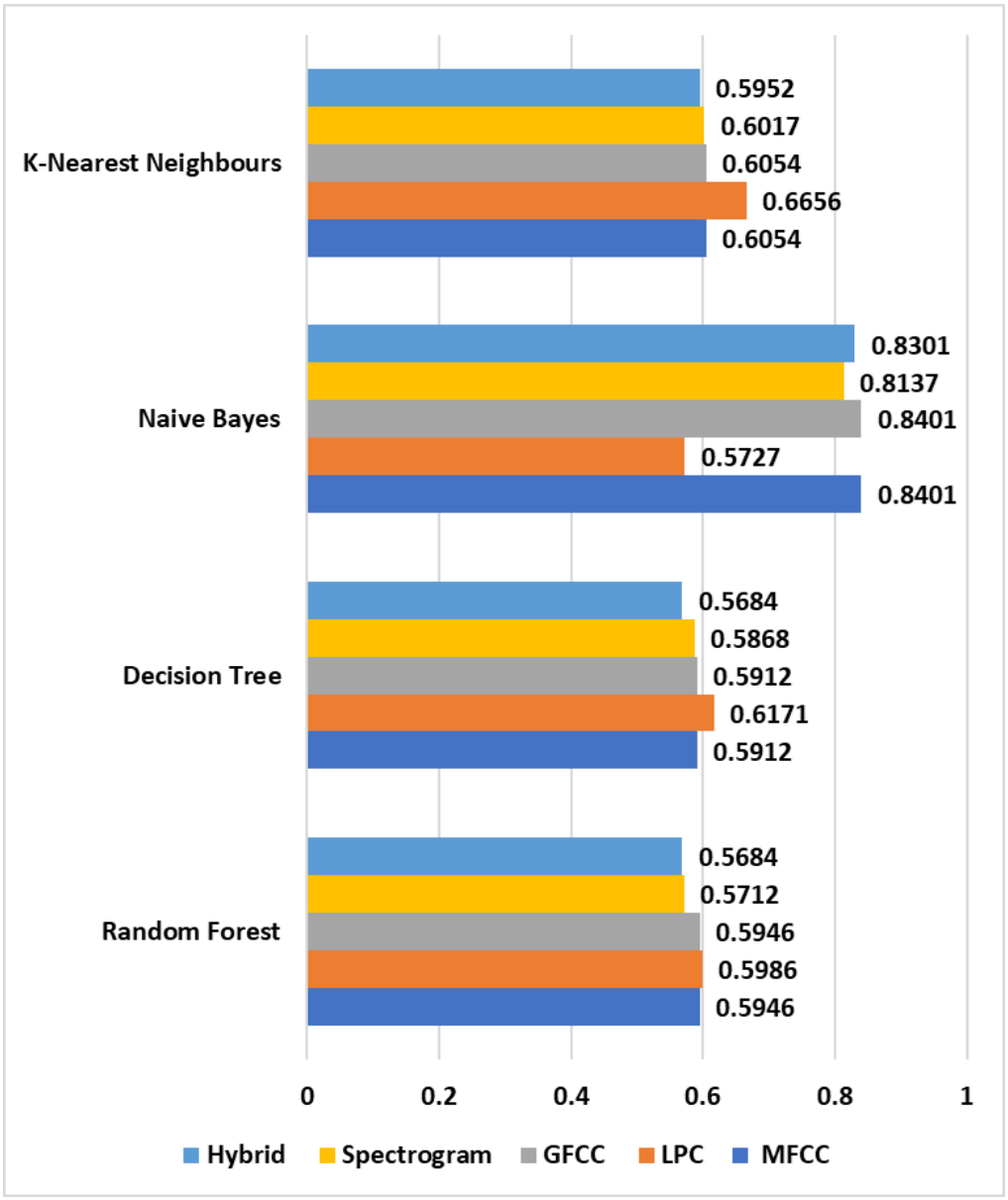

Figure 4.3: Precision Measure for Different Classifiers using Various Feature Extraction Methods

The precision values reveal that Naive Bayes achieved the highest precision with MFCC and GFCC (0.8401), outperforming other classifiers and feature sets. K-Nearest Neighbours showed the best precision with LPC (0.6656), indicating its sensitivity to this feature. Decision Tree performed best with LPC (0.6171), while Random Forest had the lowest precision across feature sets, with the hybrid approach resulting in the least precision (0.5684). These results suggest that different classifiers are sensitive to different feature sets, with Naive Bayes benefiting most from MFCC and GFCC features.

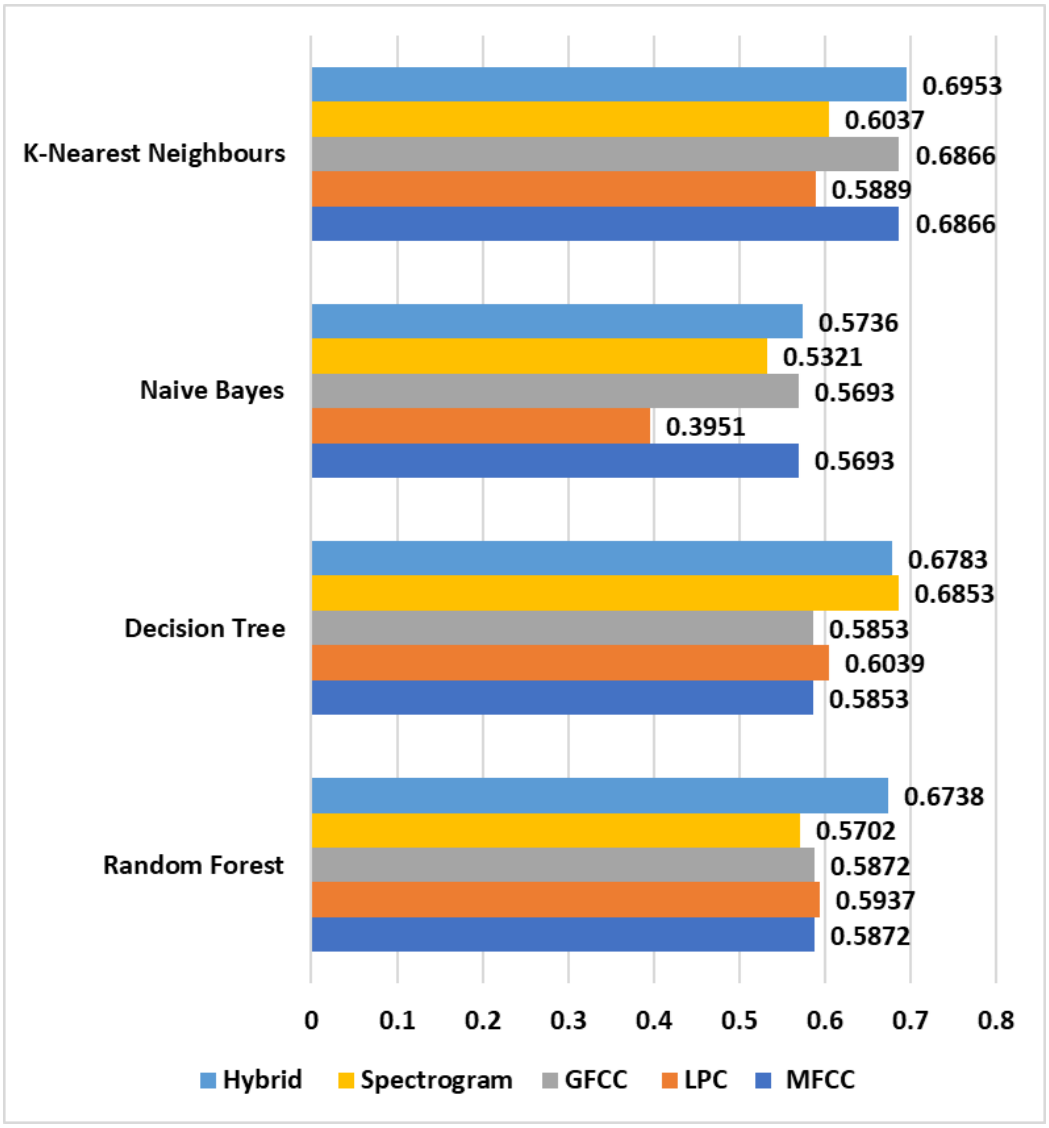

Figure 4.4: F-Measure Measure for Different Classifiers using Various Feature Extraction Methods

The F-measure results show that K-Nearest Neighbours achieved the highest F-measure with the hybrid feature set (0.6953), indicating a strong balance between precision and recall. Decision Tree also performed well with the hybrid feature set (0.6783), outperforming other feature sets. Random Forest showed an increase in F-measure with the hybrid features (0.6738), while Naive Bayes had a relatively low F-measure across all feature sets, with the hybrid set offering a slight improvement (0.5736). These findings suggest that the hybrid feature set generally improves F-measure, especially for K-Nearest Neighbours and Decision Tree.

Hypothesis Testing Results:

Rejecting the null hypothesis ($H_0 1$) indicates that the hybrid feature extraction method, which combines MFCC, LPC, GFCC, and spectrogram features, does significantly enhance classification performance compared to using individual feature sets alone. This outcome suggests that the integration of these diverse features captures more comprehensive and relevant characteristics of audio signals, leading to improved model accuracy and precision. Thus, the findings support the advantage of a hybrid approach, showing that combining multiple feature types can yield a more robust and discriminative model for audio classification tasks than relying on single-feature methods.

5. Conclusion:

In conclusion, the hybrid feature extraction method that integrates MFCC, LPC, GFCC, and Spectrogram features significantly enhances the performance of audio classification models by capturing a broader range of auditory characteristics compared to individual feature sets. The comparative analysis of classifiers, including Random Forest, Decision Tree, Naive Bayes, and K-Nearest Neighbors, revealed that combining multiple features improves classification accuracy across various metrics such as precision, recall, and F-measure. This approach addresses the limitations of single-feature methods, offering more robust performance, particularly in tasks like speaker diarization and speech recognition. The findings highlight the potential of hybrid feature extraction for real-world applications, suggesting that future work could optimize these models further by exploring additional techniques and refining algorithms for improved accuracy and reliability in diverse acoustic environments.

References:

- Abdusalomov, A. B., Safarov, F., Rakhimov, M., Turaev, B., & Whangbo, T. K. (2022). Improved feature parameter extraction from speech signals using machine learning algorithm. Sensors, 22(21), 8122. https://doi.org/10.3390/s22218122
- Ahmed, A. I., Chiverton, J. P., Ndzi, D. L., & Al-Faris, M. M. (2022). Channel and channel subband selection for speaker diarization. Computer Speech & Language, 75, 101367. https://doi.org/10.1016/j.csl.2022.101367
- Jaffino, G., Raman, R., & Jose, J. P. (2021). Improved speaker identification system based on MFCC and DMFCC feature extraction technique. In Proceedings of the 2021 Fourth International Conference on Electrical, Computer and Communication Technologies (ICECCT), Erode, India, 1-5. https://doi.org/10.1109/ICECCT52121.2021.9616805
- Coria, J. M., Bredin, H., Ghannay, S., & Rosset, S. (2021). Overlap-aware low-latency online speaker diarization based on end-to-end local segmentation. In Proceedings of the 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Cartagena, Colombia, 1139-1146. https://doi.org/10.1109/ASRU51503.2021.9688044
- Ahmad, R., Zubair, S., & Alquhayz, H. (2020). Speech enhancement for multimodal speaker diarization system. IEEE Access, 8, 126671-126680. https://doi.org /10.1109/ ACCESS.2020.3004015
- Al-Hadithy, T. M., & Frikha, M. (2023). A real-time speaker diarization system based on convolutional neural networks architectures. In Proceedings of the 2023 5th International Congress on Human-Computer Interaction, Optimization, and Robotic Applications (HORA), Istanbul, Turkey, 1-9. https://doi.org/10.1109/HORA58378.2023.10156741
- Takashima, Y., Fujita, Y., Watanabe, S., Horiguchi, S., García, P., & Nagamatsu, K. (2021). End-to-end speaker diarization conditioned on speech activity and overlap detection. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 849-856. https://doi.org/10.1109/SLT48900.2021.9383555.