I. INTRODUCTION

The rise of audio deepfakes synthetically generated or ma nipulated speech that mimics real human voices has become a significant issue in digital security. With the development of technologies like WaveNet [1] and Generative Adversarial Networks (GANs) [2], the ability to create highly realistic syn thetic audio has increased, making it difficult to differentiate between genuine and fabricated speech. This presents major risks, including identity theft, misinformation, and fraud, un derscoring the need for effective detection systems to protect against these threats.

One of the promising approaches to detecting audio deep fakes is the use of Mel Spectrograms, which provide a time frequency representation of audio that captures both temporal and frequency information. Zheng et al. [3] demonstrated the effectiveness of Convolutional Neural Networks (CNNs) for analyzing Mel Spectrograms, achieving an accuracy of 94.5% on the ASVspoof dataset. While this approach showed great potential, it still faced challenges when applied to a broader range of audio manipulations, particularly in conditions that deviate from the controlled datasets used for training.

To improve the robustness of these models, researchers have incorporated advanced preprocessing techniques such as normalization and resizing. Gupta et al. [4] used these methods alongside CNNs, achieving an accuracy of 92%. Although these enhancements improved the model's performance, they still fell short in addressing the diversity of audio manipu lations encountered in real-world conditions. The difficulty in handling varied types of manipulation suggests that more com prehensive strategies are needed to ensure effective detection in diverse settings.

Lightweight models like MobileNetV2 have also been ex plored, especially for real-time applications where compu tational efficiency is a priority. MobileNetV2 achieved an accuracy of 90% in audio deepfake detection [5]. However, the model's compact architecture may not be capable of extracting the subtle, intricate features needed to accurately identify so phisticated manipulations. This limitation highlights the trade off between efficiency and accuracy, as MobileNetV2, while fast, may not always offer the precision required for high f idelity deepfake detection.

ResNet50, another popular model, has shown promise in various tasks due to its deep residual learning capabilities. In deepfake detection, ResNet50 achieved an accuracy of 87% [14]. However, when tested on more diverse datasets, the model faced issues such as overfitting and instability. These challenges make it difficult for ResNet50 to generalize effec tively to new, unseen data, especially in complex scenarios where the audio quality or manipulation type differs from those seen during training.

The studies discussed above highlight some of the chal lenges faced by existing models. While many perform well under controlled conditions, their performance diminishes when exposed to new or more complex types of manipulations. The models also struggle to handle the inherent variability and noise found in real-world audio, which often results in less reliable performance. Furthermore, the preprocessing techniques used in these models have not been fully optimized, and there remains an opportunity to enhance their effectiveness in improving model robustness.

Despite the promising results of models like CNNs, ResNet50, and MobileNetV2, advanced architectures like Xception have not been widely explored for audio deepfake detection. Xception, known for its ability to extract fine grained features through depthwise separable convolutions, has shown great success in image classification tasks [7]. While its potential for audio detection remains largely untapped, itrepresents an opportunity to improve both the accuracy and ef f iciency of deepfake detection systems, offering a more robust solution to the challenges identified in current approaches.

This study proposes a novel end-to-end framework that inte grates Mel Spectrograms with the Xception model to address these challenges. Our approach emphasizes the optimization of preprocessing techniques, including data augmentation, trimming, and normalization, to enhance the model's ability to generalize across a variety of conditions. By leveraging the strengths of Xception's feature extraction capabilities and improving preprocessing methods, the objective is to offer a more dependable and efficient method of identifying audio deepfakes.

This study deploys a method for detecting audio deep fakes using Mel Spectrograms as input to a deep learning model. The deployment process includes a preprocessing stage where audio data is transformed into spectrograms that highlight perceptual features. The data preparation utilizes the ASVspoof 2021 dataset, enhanced with augmentation techniques to improve robustness. For model deployment, the Xception architecture is implemented to enable efficient feature extraction and classification. The system's performance is evaluated using metrics such as accuracy, precision, recall, and F1-score, confirming its effectiveness in distinguishing between authentic and fake audio.