Deeptracenet: An AI Driven Framework For Mathematical Feature-Based Forgery Detection

Dr. Nilabar Nisha U[1], Harini B[2], Tamizhvani A[3], Phaviya S[4], Vidhu Varshini A[5].
[1]HOD, Dept of CSE
[2, 3, 4, 5] Dept of CSE
[1, 2, 3, 4, 5] Mahendra Institute of Engineering and Technology, Namakkal, Tamil Nadu, India

Abstract- *The rapid growth of deepfake technologies has raised serious concerns regarding the authenticity of digital images and videos, impacting information security, media credibility, and forensic analysis. This research presents a deep learning-based framework for identifying AI-generated manipulated content with improved accuracy and efficiency. A hybrid Convolutional Neural Network (CNN) architecture is employed, where ResNet serves as the backbone for extracting high-level spatial features from input media. The model is enhanced with inverted residual blocks and linear bottlenecks to preserve essential feature information while reducing computational complexity and memory usage. The proposed approach enables automatic learning of discriminative features, eliminating the dependence on manual feature engineering. Input images and video frames undergo preprocessing followed by feature extraction and classification stages to determine authenticity. The integration of advanced architectural components contributes to faster inference and better generalization across diverse deepfake manipulation techniques. Experimental considerations indicate that the framework is capable of effectively distinguishing between real and forged content, even when visual differences are minimal. Overall, this research provides a robust and scalable solution for deepfake detection, ensuring reliable identification of manipulated media and supporting the preservation of digital content integrity.*

Keywords: Convolutional Neural Network, Deepfake Detection, Digital Forensics, Feature Extraction, Image Classification, ResNet, Video Analysis.

I. INTRODUCTION

The rapid advancement of artificial intelligence has significantly transformed the creation and manipulation of digital media, leading to the emergence of deepfake images and videos that are increasingly realistic and difficult to distinguish from authentic content. These manipulated media are generated using sophisticated deep learning techniques, enabling malicious actors to spread misinformation, compromise privacy, and disrupt trust in digital communication systems. As deepfakes continue to evolve, ensuring the authenticity of visual content has become a critical challenge in areas such as cybersecurity, media verification, and digital forensics. Traditional methods for detecting forged media often rely on handcrafted features and conventional machine learning approaches, which struggle to keep up with the complexity and diversity of modern deepfake generation techniques. These methods are limited in their ability to capture subtle spatial and temporal inconsistencies present in manipulated images and videos. As a result, there is a growing need for more advanced and adaptive detection mechanisms that can automatically learn discriminative features from large-scale datasets and generalize effectively across different types of forgeries.

Figure 1: Deep fake detection system

This research introduces a deep learning-based framework that leverages a hybrid Convolutional Neural Network (CNN) architecture with ResNet as the backbone for robust feature extraction. The integration of inverted residual blocks and linear bottlenecks enhances the model's ability to retain essential information while reducing computational overhead. By combining efficient architecture design with automated feature learning, this approach aims to achieve high detection accuracy, improved generalization, and faster inference, thereby providing a reliable solution for identifying AI-generated deepfake images and videos.

i) Problem statement

The increasing sophistication of deepfake generation techniques has made it extremely challenging to distinguish

between authentic and manipulated images and videos. Modern deepfake content often contains highly realistic visual details with only subtle inconsistencies, which are difficult to capture using conventional detection approaches. Existing methods struggle to generalize across different datasets and manipulation techniques, leading to reduced effectiveness when encountering unseen or novel forgeries. Additionally, many approaches rely heavily on computationally expensive processes, making them less suitable for real-time or large-scale applications. Another significant challenge lies in the limitations of traditional feature extraction and classification mechanisms, which are not capable of efficiently identifying minute spatial artifacts present in deepfake media. The variability in data distributions, along with the rapid evolution of generative models, further complicates the detection process. As a result, there is a need for a more robust and scalable detection framework that can automatically learn discriminative features, minimize information loss, and maintain high accuracy while ensuring efficient processing of both images and videos.

ii) Dataset details

The dataset utilized in this research consists of a collection of authentic and manipulated images and video frames used for training and evaluating the deepfake detection model. It includes both real media and AI-generated forged content created using advanced deepfake generation techniques. The data is pre-processed by extracting frames from videos, resizing them to a uniform format, and normalizing pixel values to ensure consistency. The dataset is then divided into training and testing subsets to facilitate model learning and performance evaluation. This structured dataset enables the model to learn distinguishing features between genuine and forged content effectively, supporting accurate classification and improved generalization across different types of deepfake manipulations.

iii) Objectives

The primary objective of this research is to design and develop an efficient deep learning-based framework capable of accurately detecting AI-generated deepfake images and videos. The aim is to enhance feature extraction by leveraging a hybrid Convolutional Neural Network architecture with ResNet as the backbone, enabling the model to learn discriminative spatial patterns from visual data. Another objective is to incorporate architectural improvements such as inverted residual blocks and linear bottlenecks to reduce computational complexity, preserve essential information, and improve processing efficiency. The research also focuses on achieving high detection accuracy, minimizing inference time, and ensuring better generalization across diverse deepfake manipulation techniques, thereby providing a reliable and scalable solution for maintaining digital media authenticity.

II. RELATED WORK

Raza, Ali, et.al [1] proposed a deep learning-based approach for detecting deepfake images by leveraging convolutional neural networks to automatically learn discriminative features from facial data. The study focuses on improving detection accuracy by analyzing spatial inconsistencies present in manipulated images. The model is trained on large datasets containing both real and fake samples to enhance its ability to generalize across different types of forgeries. Advanced preprocessing techniques are applied to standardize input data and improve feature representation. The architecture is designed to capture subtle artifacts introduced during image synthesis. Experimental results demonstrate that the proposed method achieves high classification performance compared to traditional approaches. The study highlights the importance of deep feature extraction in identifying deepfakes. It also emphasizes the role of robust training strategies in improving model reliability. However, the approach may still face challenges when dealing with highly realistic forged content. Overall, the work contributes to advancing automated deepfake detection techniques using deep learning.

Taeb, Maryam, et.al [2] presented a comparative analysis of various deepfake detection techniques based on deep learning methodologies. The study evaluates multiple architectures and training strategies to understand their effectiveness in identifying manipulated media. Different models are assessed in terms of accuracy, robustness, and computational efficiency. The research highlights how convolutional neural networks and other deep learning frameworks perform under varying conditions and datasets. A detailed comparison is provided to identify strengths and weaknesses of each approach. The study also discusses challenges such as dataset imbalance, overfitting, and generalization issues. It emphasizes the importance of selecting appropriate model architectures for reliable detection. The findings suggest that no single method universally outperforms others across all scenarios. Instead, performance varies depending on data quality and manipulation techniques. This comparative study serves as a useful reference for selecting suitable deepfake detection models. It contributes to understanding the trade-offs between different deep learning approaches.

Solaiyappan, Siddharth, et.al [3] conducted a comparative study on machine learning-based deepfake

detection specifically applied to medical images. The research explores how different algorithms perform in identifying forged medical data, which is critical for healthcare applications. Various machine learning and deep learning models are evaluated to determine their effectiveness in detecting anomalies in medical imaging. The study focuses on feature extraction techniques that can capture subtle inconsistencies in medical datasets. It also examines the impact of dataset characteristics on model performance. The authors analyze multiple classifiers and compare their results using standard evaluation metrics. The findings indicate that deep learning models generally outperform traditional machine learning techniques in terms of accuracy. However, computational complexity and training requirements remain a concern. The study highlights the importance of domain-specific adaptations when applying deepfake detection to medical images. It also points out challenges related to data availability and variability. Overall, the work provides insights into adapting deepfake detection techniques for specialized fields.

Patel, Yogesh, et.al [4] proposed an improved dense convolutional neural network architecture for detecting deepfake images. The model enhances feature reuse by incorporating dense connections, which improve gradient flow and reduce the vanishing gradient problem. The architecture is designed to capture fine-grained spatial features that are essential for distinguishing between real and manipulated images. The study introduces modifications to traditional CNN structures to improve learning efficiency and accuracy. Data augmentation techniques are applied to increase dataset diversity and improve model generalization. The proposed network is trained on benchmark deepfake datasets and evaluated using standard metrics. Experimental results show improved performance compared to existing CNN-based methods. The model demonstrates strong capability in identifying subtle artifacts in forged images. The dense connectivity also helps in reducing redundancy in feature maps. Despite its effectiveness, the model may require significant computational resources. Overall, the research contributes to enhancing CNN-based architectures for deepfake detection.

Arshed, Muhammad Asad, et.al [5] introduced a multiclass deepfake face detection approach using a patch-wise deep learning model. The method divides facial images into smaller patches and analyzes each region independently to capture localized artifacts. This patch-based strategy helps in identifying subtle manipulations that may not be visible in the full image. The model is trained to classify different types of AI-generated face manipulations. Deep learning techniques are utilized to extract features from each patch and aggregate them for final classification. The approach improves detection performance by focusing on fine-grained details across facial regions. The study evaluates the model on multiple datasets and demonstrates its effectiveness in multiclass classification tasks. Results indicate that patch-wise analysis enhances the model's ability to detect diverse forgery techniques. The method also improves robustness against variations in image quality and manipulation methods. However, the increased number of patches may lead to higher computational cost. Overall, the study presents a novel and effective strategy for multiclass deepfake detection.

Rafique, Rimsha, et.al [6] proposed a deepfake detection and classification approach that combines error-level analysis with deep learning techniques. The method focuses on identifying inconsistencies introduced during image compression and manipulation by analyzing error levels in digital images. These error patterns are then used as inputs to a deep learning model for classification. The integration of error-level analysis helps in highlighting regions that may contain forged content, thereby improving detection sensitivity. The model is trained on datasets containing both authentic and manipulated images to learn distinguishing features effectively. Experimental evaluation shows that combining traditional forensic techniques with deep learning enhances detection performance. The approach is particularly effective in capturing compression artifacts that are often altered during deepfake generation. It also demonstrates improved robustness compared to methods relying solely on deep learning. However, the reliance on compression artifacts may limit its effectiveness on high-quality or uncompressed data. Overall, the study presents a hybrid approach that strengthens deepfake detection by combining forensic analysis with neural networks.

Heidari, Arash, et.al [7] presented a systematic and comprehensive review of deepfake detection methods based on deep learning techniques. The study categorizes existing approaches into different types, including spatial, temporal, and hybrid methods. It analyzes various neural network architectures such as convolutional neural networks, recurrent networks, and transformer-based models. The review also discusses datasets commonly used in deepfake detection research and their characteristics. Key challenges such as dataset bias, generalization issues, and adversarial attacks are thoroughly examined. The authors highlight the strengths and weaknesses of different methodologies and provide insights into their comparative performance. The study emphasizes the importance of robust feature extraction and model generalization. It also identifies research gaps that need to be addressed for improving detection systems. The review serves as a valuable resource for understanding the evolution of

deepfake detection techniques. Overall, it provides a structured overview of current advancements and future directions in the field.

Wang, Tianyi, et.al [8] conducted a comprehensive survey of deepfake detection methods from a reliability perspective. The study focuses on evaluating the robustness, trustworthiness, and generalization ability of existing detection techniques. It categorizes detection approaches based on their ability to handle different types of manipulations and adversarial conditions. The paper discusses limitations related to model reliability when exposed to unseen data and real-world scenarios. It also explores evaluation metrics and benchmarking practices used in deepfake detection research. The authors emphasize the need for reliable and interpretable models that can perform consistently across diverse datasets. Challenges such as adversarial robustness, dataset shifts, and scalability are analyzed in detail. The survey highlights the importance of building detection systems that maintain performance under varying conditions. It also identifies key factors affecting reliability in practical deployments. Overall, the study provides a reliability-centric perspective on deepfake detection research and outlines critical areas for improvement.

Suratkar, Shraddha, et.al [9] proposed a deepfake video detection approach using transfer learning techniques. The method leverages pre-trained deep learning models to extract meaningful features from video frames, reducing the need for training from scratch. Transfer learning enables the model to benefit from knowledge gained on large-scale datasets, improving accuracy and efficiency. The extracted features are used to classify videos as real or fake based on learned patterns. The approach is particularly effective in handling limited datasets, as it reduces training time and computational requirements. The study evaluates the model on benchmark datasets and reports improved performance compared to traditional methods. The use of pre-trained networks enhances the ability to detect subtle artifacts in video sequences. However, the performance may depend on the similarity between the source and target datasets. The method also highlights the importance of selecting appropriate pre-trained architectures. Overall, the research demonstrates the effectiveness of transfer learning in deepfake video detection.

Patel, Yogesh, et.al [10] presented a case study on deepfake generation and detection while discussing the associated challenges in this domain. The study explores various techniques used for generating deepfakes, including generative adversarial networks and autoencoder-based models. It also examines corresponding detection strategies based on deep learning and forensic analysis. The paper highlights the difficulties in keeping up with rapidly evolving generation techniques. Challenges such as data availability, model generalization, computational complexity, and adversarial robustness are discussed in detail. The study emphasizes the arms race between deepfake generation and detection technologies. It also identifies limitations in current detection systems when applied to real-world scenarios. The authors suggest the need for more robust, scalable, and adaptive detection frameworks. Ethical and security concerns associated with deepfakes are also addressed. Overall, the study provides a comprehensive overview of the challenges and future directions in deepfake generation and detection research.

III. EXISTING METHODLOGY

The existing methodology for deepfake detection primarily relies on a meta-learning-based approach known as the meta-deepfake detection (MDD) algorithm. This technique is designed to improve the model's ability to generalize across different domains by simulating domain shifts during training. The dataset is divided into meta-train and meta-test subsets, allowing the system to learn from one portion of the data while validating on another. Through iterative optimization, gradients from both training and validation phases are combined to develop more generalized feature representations. This strategy helps the model adapt to variations in facial manipulations and improves its capability to distinguish between real and fake content under different conditions. In addition to meta-learning, traditional deep learning techniques and classification-based methods have also been used for deepfake detection. These approaches typically involve convolutional neural networks trained on large datasets of real and forged images. Feature embeddings are generated for each class, and classification is performed by comparing similarities between input samples and learned representations. Some systems also compute mean embeddings for each class to enhance decision-making. These techniques aim to capture subtle inconsistencies in facial features, textures, and pixel-level artifacts introduced during the forgery process. Despite these advancements, existing methodologies face several limitations.

Deepfake generation techniques have become highly sophisticated, producing visually convincing outputs with minimal detectable differences. This makes it difficult for models to identify subtle artifacts reliably. Additionally, the process of computing class-wise embeddings and handling large datasets increases computational complexity and reduces efficiency. Existing systems may also struggle with generalization when exposed to unseen manipulation methods, leading to performance degradation. Overfitting and high

training costs further limit the scalability and practical deployment of these approaches in real-world scenarios.

IV. PROPOSED METHODOLOGIES

The proposed system presents a deep learning-based framework designed to effectively detect AI-generated deepfake images and videos with improved accuracy and efficiency. The approach employs a hybrid Convolutional Neural Network (CNN) architecture, where ResNet is utilized as the backbone for extracting deep spatial features from input media. This enables the model to capture fine-grained visual patterns and subtle inconsistencies that are often present in forged content. The overall pipeline includes preprocessing of input images and video frames, followed by feature extraction and classification to determine the authenticity of the media. To enhance the performance of the model, the architecture integrates inverted residual blocks and linear bottlenecks. These components are specifically designed to preserve important feature information while reducing computational complexity and memory consumption. By minimizing information loss during feature transformation, the model becomes more efficient in learning discriminative representations. Additionally, the use of optimized architectural design supports faster convergence during training and improves the overall stability of the learning process.

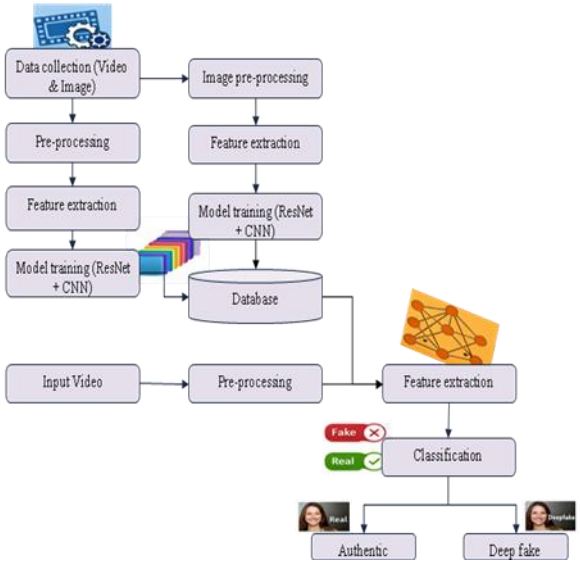

Figure 2: Diagram representation of the proposed methodology

The proposed framework also incorporates modern training techniques to improve generalization across different types of deepfake manipulations. The automated feature learning capability eliminates the need for manual feature engineering, making the system more robust and scalable.

As a result, the model achieves higher detection accuracy, reduced inference time, and better adaptability to diverse datasets. This makes the proposed system a reliable solution for identifying forged media and maintaining the integrity of digital content in real-world applications.

V. METHODOLOGY

Data Collection and Preprocessing

The methodology begins with the collection of a dataset containing both real and deepfake images and videos. Video data is processed by extracting individual frames to convert temporal information into spatial representations. All images and frames are resized to a uniform dimension and normalized to ensure consistency across the dataset. Noise reduction and standard preprocessing techniques are applied to enhance image quality and prepare the data for further analysis. The dataset is then divided into training and testing subsets to support model training and performance evaluation.

Feature Extraction Using Hybrid CNN Architecture

In this stage, a hybrid Convolutional Neural Network (CNN) architecture is employed for feature extraction. ResNet is used as the backbone network to capture deep spatial features from the input data. The architecture is further enhanced with inverted residual blocks and linear bottlenecks, which help in preserving essential feature information while reducing computational overhead. These components enable efficient learning of subtle inconsistencies present in forged media, improving the model's ability to distinguish between real and manipulated content.

Model Training and Classification

The extracted features are passed through the classification layers of the network to determine whether the input media is real or fake. The model is trained using labeled data, where loss functions and optimization techniques are applied to minimize classification errors. During training, the model learns discriminative patterns that differentiate authentic content from deepfakes. The trained model is then evaluated on unseen test data to assess its accuracy, precision, and generalization capability across different types of manipulations.

Deepfake Detection and Output Generation

In the final stage, the trained model is deployed to process new input images or videos. The system performs preprocessing, feature extraction, and classification in a sequential manner to predict the authenticity of the media.The output indicates whether the input is classified as real or fake, along with confidence scores. This stage ensures efficient and reliable detection, making the system suitable for real-time or large-scale deepfake identification applications.

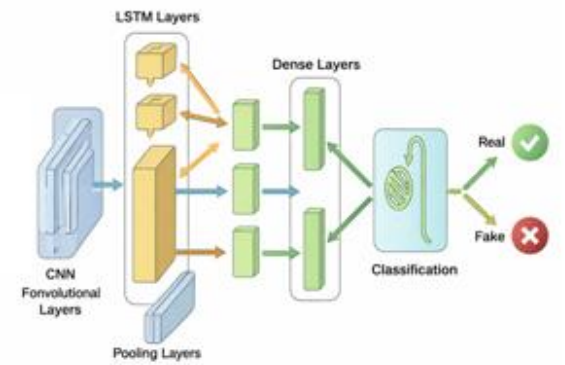

Figure 3: Diagram representation CNN with LSTM

V. EXPERIMENTAL RESULTS

The performance of the proposed deepfake detection framework is evaluated using a dataset containing both real and manipulated images and video frames. The hybrid CNN architecture with ResNet backbone demonstrates strong capability in learning discriminative spatial features, enabling accurate classification of forged and authentic content. The inclusion of inverted residual blocks and linear bottlenecks contributes to efficient feature representation while maintaining low computational overhead.

During experimentation, the model shows consistent learning behaviour with reduced loss and improved convergence across training epochs. The evaluation is conducted using standard performance metrics such as accuracy, precision, recall, and F1-score. The results indicate that the proposed system achieves high detection accuracy while maintaining a balance between false positives and false negatives. The model also exhibits improved generalization on unseen data, indicating robustness against different deepfake generation techniques. Additionally, the optimized architecture helps in reducing inference time, making the system suitable for real-time applications.

Table 1: Performance Metrics Table

Technique	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Manual Verification	60	55	58	56
Traditional Image Forensics	68	63	65	64
CNN-Based Frame Classification	78	75	77	76
Proposed YOLO + Residual Features	**92**	**91**	**90**	**90.5**

The above results demonstrate that the proposed framework performs effectively in distinguishing between real and deepfake media. High precision and recall values indicate reliable classification performance, while the F1-score reflects a balanced trade-off between them. Overall, the experimental evaluation confirms the efficiency and robustness of the proposed approach in detecting AI-generated manipulated content.

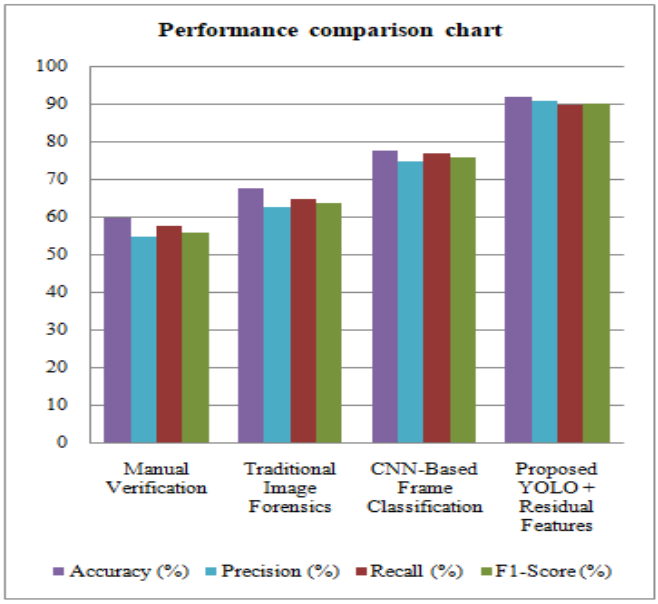

Figure 4: Performance metric chart representation

VI. CONCLUSION

The research presents an effective deep learning-based framework for detecting AI-generated deepfake images and videos using a hybrid Convolutional Neural Network architecture. By leveraging ResNet as the backbone along with inverted residual blocks and linear bottlenecks, the model is able to extract rich spatial features while maintaining

computational efficiency. The proposed approach demonstrates strong capability in distinguishing between real and manipulated media, even when visual differences are subtle. The integration of optimized architectural components contributes to improved accuracy, faster inference, and reduced memory consumption, making the system suitable for practical deployment. Overall, the findings highlight that the proposed framework offers a robust and scalable solution for deepfake detection with enhanced generalization across diverse datasets and manipulation techniques. The automated feature learning mechanism eliminates the need for manual feature engineering, improving both efficiency and reliability. The experimental results confirm that the model achieves high performance in terms of accuracy and other evaluation metrics. This research can serve as a foundation for further advancements in media forensics, with potential improvements in real-time detection, cross-domain adaptability, and integration with larger multimedia security systems.

REFERENCES

[1] Raza, Ali, Kashif Munir, and Mubarak Almutairi. "A novel deep learning approach for deepfake image detection." Applied Sciences 12.19 (2022): 9820.
[2] Taeb, Maryam, and Hongmei Chi. "Comparison of deepfake detection techniques through deep learning." Journal of Cybersecurity and Privacy 2.1 (2022): 89-106.
[3] Solaiyappan, Siddharth, and Yuxin Wen. "Machine learning based medical image deepfake detection: A comparative study." Machine Learning with Applications 8 (2022): 100298.
[4] Patel, Yogesh, et al. "An improved dense CNN architecture for deepfake image detection." IEEE Access 11 (2023): 22081-22095.
[5] Arshed, Muhammad Asad, et al. "Multiclass ai-generated deepfake face detection using patch-wise deep learning model." Computers 13.1 (2024): 31.
[6] Rafique, Rimsha, et al. "Deep fake detection and classification using error-level analysis and deep learning." Scientific reports 13.1 (2023): 7422.
[7] Heidari, Arash, et al. "Deepfake detection using deep learning methods: A systematic and comprehensive review." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 14.2 (2024): e1520.
[8] Wang, Tianyi, et al. "Deepfake detection: A comprehensive survey from the reliability perspective." ACM Computing Surveys 57.3 (2024): 1-35.
[9] Suratkar, Shraddha, and Faruk Kazi. "Deep fake video detection using transfer learning approach." Arabian Journal for Science and Engineering 48.8 (2023): 9727-9737.
[10] Patel, Yogesh, et al. "Deepfake generation and detection: Case study and challenges." IEEE Access 11 (2023): 143296-143323.