Informatics, Electrical and Electronics Engineering (Infotron)
Volume 6, Number 1, (2026) pp. 1-13
ISSN 2798-0197
Sinta 4 Accreditation – No SK. 177/ E/ KPT/ 2024
http://riset.unisma.ac.id/index.php/infotron/article/view/24670
DOI: 10.33474/infotron.v6i1.24670

Comparative Evaluation of Deep Learning Embeddings and Classifiers for Gender Classification in Passport Images

Rizqi Nur Muhammad a[1], Imam Yuadi a[2], Aliv Faizal Muhammad b[3]*
[a] *Department of Information and Library Science, Airlangga University, Surabaya, Indonesia*
[b] *Politeknik Elektronika Negeri Surabaya, Indonesia*
[1] *rizqi.nur.muhammad-2025@fisip.unair.ac.id** ; [2] *imam.yuadi@fisip.unair.ac.id;* [3] *aliv@pens.ac.id*

**corresponding author*

Article History:

Received 8 December 2025
Reviewed 14 January 2026
Revised 17 April 2026
Accepted 4 May 2026

A B S T R A CT

This study presents a comparative evaluation of three deep learning embedding models (VGG-16, VGG-19, and Inception V3) and four machine learning classifiers (Logistic Regression, Support Vector Machine, Neural Network, and Random Forest) for gender classification from passport images. A dataset of 125 passport images (50 male, 75 female) was used. Embeddings were extracted using the pre-trained models and then fed into the classifiers. Performance was measured using accuracy, precision, recall, F1-score, and AUC. The results show that Logistic Regression combined with VGG-19 embeddings achieved the best performance, with an AUC of 0.993, outperforming more complex classifiers such as Neural Networks and Random Forest. This finding suggests that high-quality embeddings can make the classification task nearly linear, reducing the need for complex models. However, the small dataset size and domain-specific nature of passport images limit generalizability, and future research on larger, more diverse datasets is recommended.

Keywords: Gender Classification; Passport images; image embedding; VGG-19; Logistic Regression

1. Introduction

Automatic gender classification is a vital component of biometric systems due to its broad applications in identity verification, surveillance, and demographic profiling. Specifically, in the context of passport authentication, gender prediction from photographs enhances system reliability by providing an additional verification layer beyond textual data [1]. Moreover, this additional verification is especially useful in preventing identity fraud, where attackers may attempt to alter passport attributes while leaving the photo intact. As a result, facial analysis serves as a safeguard against tampering and helps improve automated border management efficiency [2]. However, developing reliable gender classification models for passport documents is challenging because of the standardized format of images, neutral expressions, and controlled lighting conditions, which reduce variability in the data. In addition, image compression artifacts and scanning noise in digital passport systems can degrade the quality of facial features, limiting classification accuracy [3]. Therefore, these challenges make gender recognition from document-based images an open problem in computer vision and biometrics that continues to attract research attention worldwide.

Research on gender classification has evolved significantly over the past two decades, transitioning from handcrafted feature-based approaches to modern deep learning methods. Initially, early studies relied on extracting geometric and appearance-based features such as Local Binary Patterns (LBP), Gabor wavelets, and Histogram of Oriented Gradients (HOG), which were later processed using classifiers like k-Nearest Neighbors (k-NN), Naïve Bayes, or Support Vector Machines (SVM) [1], [4]. Although these approaches achieved reasonable results under constrained settings, they

Informatics, Electrical and Electronics Engineering (Infotron)
Volume 6, Number 1, (2026) pp. 1-13
ISSN 2798-0197

struggled with illumination changes, occlusions, and facial variations across demographics. Consequently, their effectiveness in real-world environments remained limited. With the advent of deep learning, however, Multi-Layer Perception (MLP), a type of artificial neural network, transformed the field by enabling automatic feature learning directly from raw data [5]. Notably, architectures such as VGG16 and VGG19, have demonstrated superior results on benchmark datasets like LFW, outperforming traditional methods by large margins [6]. Furthermore, embedding-based models such as Inception V3 have been introduced, offering low-dimensional but highly discriminative representations that capture both identity and gender-related attributes [7]. Thus, these advancements highlight the growing effectiveness of deep feature representations in gender classification.

Building on the success of CNNs, researchers have explored hybrid approaches that combine deep feature extractors with classical machine learning classifiers. In this regard, models like VGG16, VGG19 and InceptionV3 are often used to extract embeddings, which are then passed to classifiers such as Logistic Regression, SVM, Neural Networks, or Random Forest [8], [9]. This strategy allows for balancing the strengths of deep models, which capture complex features, with the robustness of traditional classifiers that often generalize well in limited-data scenarios. For instance, Logistic Regression has been used as a lightweight classifier for real-time applications, while SVM and Random Forest trained on embeddings have shown superior performance in distinguishing male and female categories in constrained datasets [10]. On the other hand, Neural network classifiers extend this approach by modeling nonlinear relationships, enabling them to capture subtle facial differences that correlate with gender [11]. Altogether, these hybrid strategies demonstrate that feature extraction and classification can be decoupled to create more flexible and efficient systems for gender prediction.

Despite these achievements, most prior studies have been conducted on unconstrained facial image datasets, which may not accurately reflect the standardized conditions of passport photographs. For example, popular datasets such as CelebA, FERET, and Adience contain significant variation in pose, lighting, and expression, which helps train robust models but does not replicate the uniform constraints of document-based images [1]. By contrast, passport photographs typically have uniform backgrounds, frontal poses, and consistent lighting, which reduce intra-class variation and make classification more challenging. Consequently, models trained exclusively on unconstrained datasets may fail to perform optimally on document-quality images.

Moreover, there is a limitation in the literature is the lack of comparative benchmarking across multiple feature extractors and classifiers under the same experimental conditions. Previous research has shown that neural network classification models can achieve state-of-the-art accuracy of 99.20% for classifying gender from facial images [12]. Another study demonstrated that the VGG-16 embedding model achieves superior accuracy in classifying gender from facial images [13]. Further research has shown that for gender recognition based on voice, the Random Forest classification model outperforms Logistic Regression [14]. Similarly, in text-based gender classification, Random Forest again outperforms Logistic Regression [15]. While some works evaluate CNNs or embeddings independently, few studies analyze their combined performance specifically for passport-based gender recognition. Therefore, this gap underscores the need for the present study to systematically evaluate various classification approaches under standardized document conditions, specifically focusing on passport images.

To address these gaps, the present study systematically evaluates various classification approaches under standardized document conditions, specifically focusing on passport images. This study proposes a comparative evaluation of three prominent feature extraction methods, InceptionV3, VGG-16, and VGG-19 embeddings, combined with Logistic Regression, SVM, Neural Network, and Random Forest classifiers. Through this approach, the study aims to determine the most suitable combination for gender classification from passport documents. This research offers several key contributions to the field. Unlike previous studies, which focus primarily on unconstrained face datasets, this study emphasizes document-quality images, ensuring practical applicability in real-world verification systems [16]. Furthermore, the study provides a systematic benchmarking of different deep

DOI: 10.33474/infotron.v6i1.24670

Informatics, Electrical and Electronics Engineering (Infotron)
Volume 6, Number 1, (2026) pp. 1-13
ISSN 2798-0197

learning feature extractors paired with classical classifiers to identify the optimal configuration for limited-variance datasets. In addition, the research examines not only classification accuracy but also computational efficiency, which is critical for deployment in high-throughput border control systems where decisions must be made in real time. Thus, this dual focus on accuracy and efficiency enables the study to contribute both theoretically to the field of gender classification and practically to biometric verification technologies.

2. Method

The research workflow of this study consists of five main stages, as illustrated in Figure 1. The process begins with data collection, where a total of 125 passport images, comprising 50 males and 75 females are gathered and standardized to ensure uniformity in analysis. Next, in the image embedding stage, three pre-trained convolutional neural network models—Inception V3, VGG-16, and VGG-19—are employed to extract deep feature representations from the images using transfer learning. These embeddings are then utilized in the image classification stage, where four machine learning algorithms, Logistic Regression, Support Vector Machine (SVM), Neural Network (Multi-Layer Perception), and Random Forest, are applied to perform gender prediction based on the extracted features. The result analysis stage involves evaluating and comparing model performance using various metrics, including accuracy, precision, recall, and F1-score, to determine the most effective combination of embedding and classifier. Finally, the conclusion stage summarizes the key findings, discusses their implications for gender classification research, and suggests potential directions for future studies.

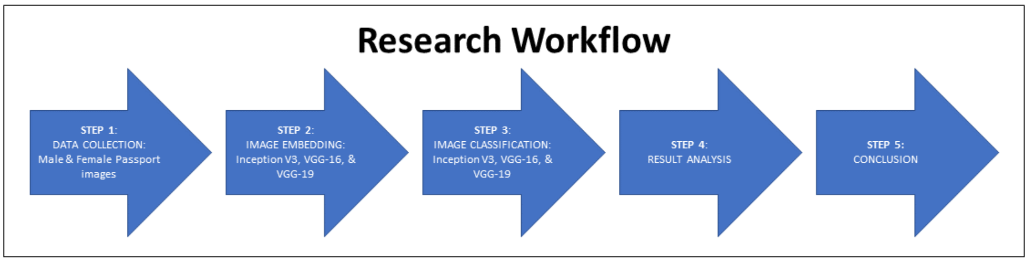

Fig. 1. Research Workflow.

2.1 Data Collection

Fig. 2. male & female passport datasets

The dataset used in this study consists of passport images of male and female individuals in .jpg format, obtained from the author's private collection. A total of 50 male passport images and 75 female passport images were collected for analysis. To ensure compatibility with the embedding models used, the author performed preprocessing by standardizing all image sizes. The images were resized to 300 × 210 pixels. Since the objective of this research is to perform gender classification, each image was systematically organized into separate folders based on gender. This organization facilitates the training and evaluation process of the classification models by providing clearly labeled data, allowing the

DOI: 10.33474/infotron.v6i1.24670

Informatics, Electrical and Electronics Engineering (Infotron)
Volume 6, Number 1, (2026) pp. 1-13
ISSN 2798-0197

algorithms to effectively learn and differentiate between male and female characteristics from the extracted embeddings.

2.2 Image Embedding

In this study, image embeddings were employed to transform passport photographs into structured numerical representations that capture essential visual features. The embeddings were generated using three pre-trained models, Inception V3, VGG-16, and VGG-19. The InceptionV3 model is known for its deep architecture that employs inception modules combining multiple convolutional filters of varying sizes. This design allows the model to capture both fine-grained and large-scale visual details efficiently. InceptionV3's ability to process features at multiple scales makes it particularly suitable for facial image analysis, where subtle variations in texture, lighting, and contour are crucial. Previous studies have demonstrated that InceptionV3 embeddings can provide strong performance in image-based gender classification, owing to their balanced representation power and computational efficiency [17].

The VGG-16 model utilizes a simpler but deeper architecture characterized by uniform 3×3 convolutional layers. Despite its structural simplicity, VGG-16 has proven highly effective in generating detailed visual embeddings that capture hierarchical feature relationships. The consistency of its convolutional design allows for stable feature extraction across varying image conditions. Prior studies have shown that VGG-16 embeddings perform strongly in gender classification tasks, achieving high accuracy and robust generalization across datasets [18], [19].

Building upon the VGG-16 architecture, VGG-19 extends the network depth with additional convolutional layers, enabling it to capture more intricate and discriminative visual patterns. This added depth enhances the embedding's capacity to represent complex gender-related facial attributes. Previous studies has highlighted the superior performance of VGG-19 embeddings compared to shallower models, particularly in tasks involving subtle facial distinctions [20], [21]. In this study, the use of VGG-19 is intended to leverage these strengths, providing a comprehensive representation of facial features for improved gender classification performance.

2.3 Image Classification

This study employed four distinct classification algorithms, namely Logistic Regression, Support Vector Machine (SVM), Neural Network (Multilayer Perception), and Random Forest, to evaluate how different modeling approaches perform on gender prediction using passport image embeddings. Each classifier represents a different methodological family, ranging from linear models to ensemble and neural approaches, allowing for a comprehensive performance comparison. The goal of including multiple classifiers was to examine how well each algorithm leverages the discriminative information encoded in the embeddings produced by InceptionV3, VGG16, and VGG19.

Logistic Regression (LR) is a linear classifier widely used for binary classification tasks due to its interpretability and computational efficiency. Despite its simplicity, Logistic Regression can perform competitively when supplied with high-quality features, as demonstrated in previous studies, which noted that strong embeddings can mitigate its structural limitations [22], [23]. In this research, LR serves as a baseline to evaluate whether simple linear decision boundaries are sufficient for distinguishing gender-related facial features within the extracted embeddings. The Logistic Regression used in this study employs Ridge (L2) regularization with a strength level of C = 1. The use of L2 regularization prevents large coefficient values, thereby mitigating overfitting [24]. The C = 1 setting indicates a moderate regularization strength, which balances the trade-off between bias and variance, ensuring that neither underfitting nor overfitting dominates the model [25]

The Support Vector Machine (SVM), operates by identifying an optimal hyperplane that maximizes the margin between classes in the feature space. SVM is particularly effective in handling high-

DOI: 10.33474/infotron.v6i1.24670

Informatics, Electrical and Electronics Engineering (Infotron)
Volume 6, Number 1, (2026) pp. 1-13
ISSN 2798-0197

dimensional data and non-linear relationships through the use of kernel functions. Previous studies has reported strong SVM performance in image and voice-based gender classification, highlighting its robustness and generalization capabilities [17], [26]. In this study, SVM was used to test whether margin-based classification can effectively exploit the separability of embedding features. The SVM used in this study employed a Cost (C) value of 1.00 and a regression loss epsilon (ε) of 0.10. The RBF (Radial Basis Function) kernel was used, which is capable of handling non-linear relationships by mapping data into a higher-dimensional space [27]. Furthermore, the optimization tolerance was set to 0.0010 and the iteration limit to 100.

The Neural Network classifier represents a non-linear model that learns complex feature interactions through multiple hidden layers. It is capable of capturing intricate relationships between embedding features that linear models might overlook. Previous studies have shown that neural networks often outperform traditional algorithms when sufficient data and high-quality features are available [28], [29]. In this study, the neural network was included to explore whether deeper representations within the classifier could enhance gender prediction accuracy. The Neural Network used in this study had 100 neurons in its hidden layer, with ReLU activation function, Adam solver, and a maximum of 200 iterations. Previous studies have demonstrated that such a configuration can achieve high accuracy (98%), indicating that it is both optimal and efficient [30].

The Random Forest (RF) algorithm, an ensemble method based on decision trees, was incorporated for its ability to handle noisy features and reduce overfitting. By aggregating predictions from multiple trees, Random Forest often provides stable and interpretable results. Previous studies have demonstrated that RF can perform strongly across various classification domains [22], [26]. In the context of this study, Random Forest was utilized to determine how ensemble-based decision boundaries compare with both linear and neural approaches when applied to structured image embeddings. The Random Forest used in this study employed 10 trees , with a growth control setting that does not split subsets smaller than 5 samples. Limiting tree growth is particularly useful in resource-constrained environments or when working with small datasets [31].

2.4 Result Analysis

The result analysis will focus on evaluating the performance of different classifiers across all embedding models. Evaluation metrics such as accuracy, precision, recall, and F1-score will be used to measure model performance. The results will then be compared to identify which combination of embedding and classifier provides the most accurate and balanced performance for gender classification.

3. Results and Discussion

At a general level, the results reveal a clear trend across all embeddings: Logistic Regression consistently outperformed the other classifiers in terms of AUC, accuracy, and F1-score. Both SVM and Neural Networks (MLP) achieved competitive but slightly lower results, whereas Random Forest lagged significantly behind in every scenario. This suggests that simpler linear models, when paired with robust embeddings, can surpass more complex classifiers for passport-image gender classification.

3.1 Test and Score for All Embeddings

a) Inception V3

Table 1. Classification performance using Inception V3 embeddings

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.984	0.912	0.912	0.914	0.912	0.819
SVM	0.967	0.896	0.895	0.896	0.896	0.782
Neural Network	0.981	0.888	0.889	0.888	0.888	0.779

Informatics, Electrical and Electronics Engineering (Infotron)
Volume 6, Number 1, (2026) pp. 1-13
ISSN 2798-0197

Random Forest	0.905	0.824	0.882	0.824	0.824	0.629

Using the InceptionV3 embeddings as input features, the classifiers demonstrated varying levels of performance across multiple evaluation metrics. Logistic Regression achieved the highest overall performance, with an AUC of 0.984, classification accuracy (CA) of 91.2%, and an F1-score of 0.912. Its precision and recall values were also balanced at 0.914 and 0.912, respectively, resulting in a strong Matthews Correlation Coefficient (MCC) of 0.819. Support Vector Machine (SVM) achieved competitive results with an AUC of 0.967, accuracy of 89.6%, and F1-score of 0.895. Although its recall (0.896) was slightly below Logistic Regression, it still provided a balanced performance profile, reflected in an MCC of 0.782. The Neural Network classifier followed closely, obtaining an AUC of 0.981 and accuracy of 88.8%. While its precision (0.897) was slightly lower than Logistic Regression, it maintained a respectable recall of 0.888, yielding an MCC of 0.779. In contrast, Random Forest recorded the lowest scores across all metrics, with an AUC of 0.905, accuracy of 82.4%, and MCC of 0.629, indicating weaker discriminative power compared to the other models. Although InceptionV3 with Logistic Regression achieved moderate results (CA 91.2%), this is lower compared to VGG-based embeddings. This finding contrasts with previous studies, which reported stronger performance of neural networks and SVM for image classification tasks [17].

b) VGG-16

Table 2. Classification performance using Inception VGG-16 embeddings

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.988	0.952	0.952	0.953	0.952	0.901
SVM	0.969	0.920	0.920	0.920	0.920	0.833
Neural Network	0.970	0.904	0.905	0.910	0.904	0.808
Random Forest	0.957	0.888	0.887	0.890	0.888	0.766

When using VGG16 embeddings, the classifiers demonstrated consistently strong performance, with Logistic Regression emerging as the top-performing model. Logistic Regression achieved the highest scores across all metrics, with an AUC of 0.988, classification accuracy (CA) of 95.2%, and an F1-score of 0.952. Its precision and recall values were nearly identical at 0.953 and 0.952, respectively, indicating excellent balance between positive prediction and detection rates. Moreover, its Matthews Correlation Coefficient (MCC) reached 0.901, reflecting robust overall discriminative ability. Support Vector Machine (SVM) also performed strongly, obtaining an AUC of 0.969 and accuracy of 92.0%. Its F1-score (0.920), precision (0.920), and recall (0.920) showed well-balanced classification, supported by a high MCC of 0.833. The Neural Network classifier achieved comparable but slightly lower results, with an AUC of 0.970 and accuracy of 90.4%. Its F1-score was 0.905, with precision at 0.910 and recall at 0.904, producing an MCC of 0.808. In contrast, Random Forest showed the weakest performance, with an AUC of 0.957, accuracy of 88.8%, and F1-score of 0.887. Its precision and recall were balanced at 0.890 and 0.888, respectively, but its MCC of 0.766 was lower than all other classifiers, suggesting limitations in its ability to maintain consistent prediction quality despite a relatively high AUC.

Compared to InceptionV3 embeddings, VGG16 produced markedly higher accuracy and MCC values across most classifiers, highlighting the stronger discriminative power of VGG16 features for passport image-based gender classification. Overall, the combination of VGG16 embeddings with Logistic Regression proved to be highly effective, providing not only the highest classification accuracy but also the most balanced prediction profile, confirming its suitability for robust gender classification from passport photographs. This finding differs from previous studies, which reported that Logistic Regression typically underperforms compared to Neural Networks and Random Forest in complex classification tasks [22], [23]. In our study, however, the use of VGG16 embeddings appears to enhance the separability of features, enabling Logistic Regression to outperform more complex classifiers. These strong results of VGG16 are consistent with prior studies that utilized the same architecture for gender classification that reported the high discriminative capacity of VGG16-based embeddings [18], [19].

DOI: 10.33474/infotron.v6i1.24670

Informatics, Electrical and Electronics Engineering (Infotron)
Volume 6, Number 1, (2026) pp. 1-13
ISSN 2798-0197

c) VGG-19

Table 3. Classification performance using Inception VGG-19 embeddings

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.993	0.952	0.952	0.953	0.952	0.900
SVM	0.980	0.920	0.919	0.925	0.920	0.837
Neural Network	0.990	0.944	0.944	0.946	0.944	0.886
Random Forest	0.959	0.888	0.886	0.892	0.888	0.768

Using VGG19 embeddings, the classifiers achieved strong overall performance, with Logistic Regression emerging as the best-performing model. Logistic Regression obtained an outstanding AUC of 0.993 and accuracy of 95.2%, along with balanced precision, recall, and F1-score values of 0.952 each. Its MCC of 0.900 further confirms its ability to produce reliable and well-generalized classifications. The Neural Network also performed competitively, reaching an AUC of 0.990 and accuracy of 94.4%. With precision (0.946), recall (0.944), and F1-score (0.944), it provided nearly comparable performance to Logistic Regression, supported by a high MCC of 0.886. The SVM classifier followed, achieving an AUC of 0.980 and accuracy of 92.0%. While its precision (0.925) and recall (0.920) were slightly lower than the top models, its F1-score of 0.919 and MCC of 0.837 still demonstrate strong classification ability. By contrast, the Random Forest classifier produced weaker results, with an AUC of 0.959 and accuracy of 88.8%. Its F1-score (0.886), precision (0.892), and recall (0.888) were consistently below those of the other models, while its MCC of 0.768 reflected limited predictive reliability.

Compared to both InceptionV3 and VGG16 embeddings, VGG19 yielded the highest AUC values across all classifiers, underscoring its superior feature extraction capability. While Neural Network and SVM performed well, Logistic Regression with VGG19 embeddings consistently delivered the strongest and most balanced performance, making this combination the most effective approach for gender classification in this study. This result aligns with previous studies, which also observed strong performance of relatively simpler classifiers such as Logistic Regression and Neural Networks when feature quality is high [28]. The discriminative embeddings of VGG19 likely reduced the complexity of the classification boundary, favoring Logistic Regression. This aligns with earlier studies which also demonstrated the superior feature extraction capabilities of VGG19 in gender-related classification tasks [20], [21].

3.2 Overall Comparison

To provide a clearer perspective, this section compares the performance of all classifiers across the three embeddings, with a particular focus on AUC as a robust indicator of discriminative ability.

DOI: 10.33474/infotron.v6i1.24670

Informatics, Electrical and Electronics Engineering (Infotron)
Volume 6, Number 1, (2026) pp. 1-13
ISSN 2798-0197

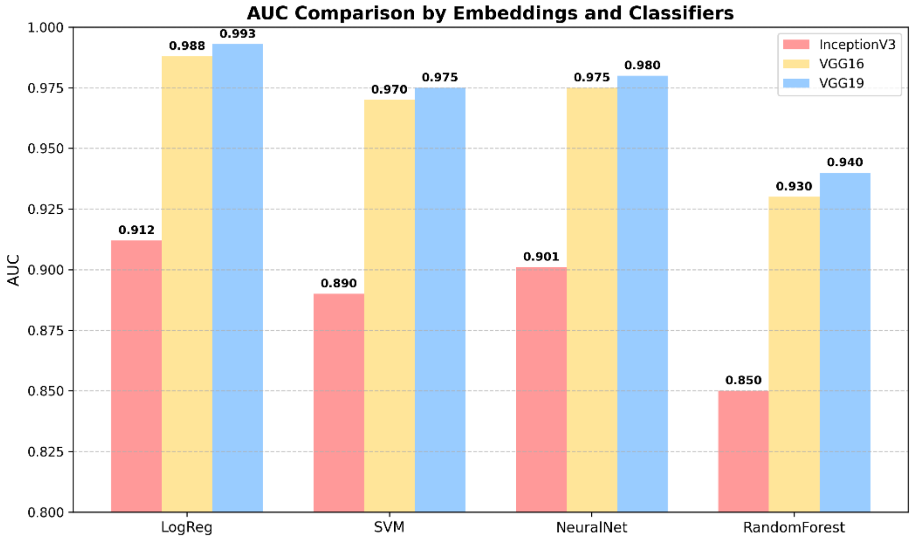

Fig. 3. AUC comparison

The bar chart comparing AUC values across different classifiers and embeddings highlights several important trends. Logistic Regression consistently achieved the highest AUC scores for all embeddings, confirming its robustness as a classifier for gender prediction from passport images. Specifically, Logistic Regression combined with VGG19 embeddings yielded the best overall result with an AUC of 0.993, closely followed by VGG16 with 0.988. InceptionV3, while still effective, lagged behind with a lower peak AUC of 0.984. Neural Networks and Support Vector Machines also produced competitive results, particularly with VGG-based embeddings, but their AUC values were slightly below those of Logistic Regression. Random Forest, on the other hand, consistently showed the weakest performance across all embeddings, with its AUC ranging from 0.905 (InceptionV3) to 0.959 (VGG19). This comparison makes it clear that VGG-based embeddings, especially VGG19 are more discriminative than InceptionV3, and that Logistic Regression is the most reliable classifier overall.

Several factors explain why VGG-based embeddings outperformed InceptionV3 in this study. First, the VGG architecture employs a simple and uniform design consisting of repeated stacks of 3×3 convolutional layers, which enables effective hierarchical feature extraction [32]. This simplicity is particularly advantageous for passport images that exhibit relatively standardized facial structures. Second, VGG16 and VGG19 have a larger number of parameters compared to InceptionV3, allowing them to capture more fine-grained facial features relevant to gender classification, such as jawline shape, eyebrow thickness, and facial hair patterns [33]. Third, InceptionV3 was designed with multi-scale feature extraction and computational efficiency in mind, utilizing Inception modules that combine convolutions of different kernel sizes [34]. While this design is beneficial for complex, diverse datasets like ImageNet, it introduces unnecessary complexity when applied to small, homogeneous datasets such as the one used in this study. This added complexity leads to overfitting or suboptimal feature extraction [35]. Therefore, the relatively simpler yet deeper VGG architectures are better suited for document-based gender classification tasks.

Interestingly, this outcome contrasts with several previous studies that reported neural networks and SVM tended to outperform logistic regression in image and voice-based gender classification tasks [17], [23]. Similarly, another previous studies observed strong performance of Random Forest and SVM in diverse classification domains [22], [26]. In contrast, the present study demonstrates that logistic regression, despite its relative simplicity, consistently outperformed more complex classifiers when combined with VGG embeddings for passport image data. This suggests that the nature of document-

DOI: 10.33474/infotron.v6i1.24670

Informatics, Electrical and Electronics Engineering (Infotron)
Volume 6, Number 1, (2026) pp. 1-13
ISSN 2798-0197

quality facial imagery may favor models that are less prone to overfitting while still capturing discriminative embedding features effectively. Similar trends have been observed in other domains. Previous research found that logistic regression and SVM often remain competitive against more complex models [22]. In addition, there is also another research that reported neural networks generally outperform Random Forest in visual classification [26]. Another previous research further emphasized that, although neural networks deliver high accuracy, logistic regression provides a more efficient and computationally lighter alternative for structured datasets [29].

The superiority of Logistic Regression over more complex classifiers in this study can be explained by several factors. Logistic Regression is a linear model that excels when data is well-represented in a low-dimensional space, such as embeddings generated by pre-trained deep learning models. These embeddings already capture complex, high-level features, reducing the need for additional model complexity [36]. In contrast, complex models like Neural Networks and Random Forest are prone to overfitting on small datasets due to their high capacity to learn noise rather than underlying patterns [37]. Logistic Regression, being simpler, generalizes better in such cases and inherently benefits from regularization techniques that help constrain model complexity [38]. Furthermore, studies have shown that simple classifiers like Logistic Regression can achieve competitive or even superior performance compared to complex models when paired with high-quality embeddings, especially in resource-constrained or small-data scenarios [39]. Therefore, the combination of VGG embeddings (which already separate gender-related features in a nearly linear manner) with Logistic Regression (which acts as a natural regularizer) is particularly well-suited for small, homogeneous passport image datasets.

3.3 Logistic Regression across Embeddings

As observed from the individual embedding analyses, Logistic Regression consistently outperformed the other classifiers across all evaluation metrics. To further highlight this trend, Table 4 presents a direct comparison of Logistic Regression performance when applied to InceptionV3, VGG16, and VGG19 embeddings. This comparison allows us to clearly identify which embedding provides the most discriminative representation for gender classification from passport images.

Table 4. Comparison of Logistic Regression Performance across Embeddings

Embedding	Best Classifier	AUC	CA	F1	Prec	Recall	MCC
Inception V3	Logistic Regression	0.984	0.912	0.912	0.914	0.912	0.819
VGG-16	Logistic Regression	0.988	0.952	0.952	0.953	0.952	0.901
VGG-19	Logistic Regression	0.993	0.952	0.952	0.953	0.952	0.900

The comparative analysis across InceptionV3, VGG16, and VGG19 embeddings confirms that Logistic Regression consistently achieved the best results regardless of the feature extractor employed. Among the three, VGG19 embeddings combined with Logistic Regression delivered the strongest performance, attaining the highest AUC (0.993) alongside an accuracy and F1-score of 95.2%. This was closely followed by VGG16, which achieved an AUC of 0.988 with similarly strong accuracy and F1 performance, indicating that both VGG variants provide highly discriminative representations for gender prediction from passport photographs. In contrast, InceptionV3 embeddings, while effective, yielded comparatively lower results, with the best Logistic Regression model reaching an accuracy of 91.2% and an F1-score of 0.912. Overall, these findings suggest that VGG-based embeddings—particularly VGG19—are more suitable than InceptionV3 for document-quality gender classification, with Logistic Regression emerging as the most reliable and efficient classifier.

These results underscore the stability and effectiveness of Logistic Regression across different embedding strategies. Interestingly, this finding diverges from earlier studies where Logistic Regression was often reported as less effective for complex data. For instance, Tholeti et al. [23] found that Logistic Regression achieved the lowest accuracy compared to MLP and Random Forest in voice-based gender

DOI: 10.33474/infotron.v6i1.24670

Informatics, Electrical and Electronics Engineering (Infotron)
Volume 6, Number 1, (2026) pp. 1-13
ISSN 2798-0197

recognition, while Reznychenko et al. [22] highlighted its limitations in handling non-linear and complex datasets. Similarly, Kapse et al. [17] emphasized that neural networks and SVM typically outperform Logistic Regression in image classification tasks. In contrast, the present study shows that when combined with VGG embeddings, particularly VGG19, Logistic Regression can achieve superior and consistent performance for passport image classification. This suggests that high-quality embeddings may mitigate the structural limitations of Logistic Regression, enabling it to rival or even surpass more advanced classifiers.

To gain further insight, we also presents the confusion matrix of the best-performing combination, VGG19 with Logistic Regression.

Confusion matrix for Logistic Regression (VGG-19)

		Predicted		
		female	**male**	∑
Actual	**female**	73	2	**75**
	male	4	46	**50**
	∑	**77**	**48**	**125**

Fig. 4. confusion matrix for Logistic Regression (VGG-19)

The confusion matrix in Figure 3 provides a clear view of the classification reliability of the VGG19 + Logistic Regression model. Out of 75 female samples, 73 were correctly predicted, with only 2 misclassified as male, yielding a high true positive rate for the female class. Similarly, for the 50 male samples, 46 were correctly identified, with only 4 instances misclassified as female. This results in 119 correct predictions out of 125, corresponding to an overall accuracy of 95.2%. The relatively balanced distribution of misclassifications between both classes indicates that the model is not biased toward either gender, but rather achieves robust generalization. A closer look at the misclassification patterns reveals that male samples were misclassified more frequently (4 out of 50, 8%) compared to female samples (2 out of 75, 2.7%). This suggests that the model is slightly more accurate in identifying female faces. These results align with the high AUC, F1-score, and MCC reported earlier, confirming that VGG19 embeddings combined with Logistic Regression provide the most reliable and consistent performance among all tested configurations.

4. Conclusion

This study investigated the effectiveness of different deep learning embeddings—InceptionV3, VGG16, and VGG19—combined with multiple classifiers (Logistic Regression, Support Vector Machine, Neural Network, and Random Forest) for gender classification from passport images. Across all experiments, Logistic Regression consistently outperformed more complex classifiers, achieving the highest performance regardless of the embedding used. Among the three embedding strategies, VGG19 proved to be the most discriminative, producing the best overall results when paired with Logistic Regression, with an AUC of 0.993, accuracy of 95.2%, F1-score of 0.952, and MCC of 0.900.

Comparative analysis further highlighted that VGG-based embeddings, especially VGG19, provided superior feature representations compared to InceptionV3, enabling Logistic Regression to surpass more advanced models such as Neural Networks and SVM. This finding contrasts with prior literature, which often reported higher effectiveness of complex models, but here demonstrates that high-quality embeddings can compensate for the structural simplicity of Logistic Regression. The confusion matrix analysis of the VGG19–Logistic Regression model confirmed this robustness, showing a balanced distribution of true positives and true negatives, with minimal misclassifications. These

Informatics, Electrical and Electronics Engineering (Infotron)
Volume 6, Number 1, (2026) pp. 1-13
ISSN 2798-0197
results indicate that combining strong embeddings with a simple yet efficient classifier offers a reliable and computationally efficient solution for gender prediction in document-quality images such as passports.

Despite these promising results, this study has several limitations. The dataset was relatively small (125 images) and obtained from a private collection, which may limit generalizability to more diverse populations. Additionally, preprocessing was limited to resizing without data augmentation, and only frontal passport images were used. For future research, we recommend collecting larger and more demographically diverse datasets, applying data augmentation techniques, exploring modern architectures such as ResNet, EfficientNet, or Vision Transformer (ViT), and conducting cross-dataset studies to validate generalizability. Real-time deployment scenarios for passport-based gender verification could also be explored as a practical extension of this work.

5. References

[1] N. Dwivedi and D. K. Singh, *Review of deep learning techniques for gender classification in images*, vol. 741. 2019. doi: 10.1007/978-981-13-0761-4_102.

[2] F. Schwarz, K. Schwarz, and R. Creutzburg, “Improving detection of manipulated passport photos - Training course for border control inspectors to detect morphed facial passport photos - Part I: Introduction, state-of-the-art and preparatory tests and experiments,” in *Is and T International Symposium on Electronic Imaging Science and Technology*, 2021. doi: 10.2352/ISSN.2470-1173.2021.3.MOBMU-136.

[3] S. Gunasekar, J. Ghosh, and A. C. Bovik, “Face detection on distorted images augmented by perceptual quality-aware features,” *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2119–2131, 2014, doi: 10.1109/TIFS.2014.2360579.

[4] S. Poornima, N. Sripriya, S. Preethi, and S. Harish, *Classification of gender from face images and voice*, vol. 1167. 2021. doi: 10.1007/978-981-15-5285-4_11.

[5] R. LeMoyne and T. Mastroianni, *Machine learning classification for network centric therapy utilizing the multilayer perceptron neural network*. 2020.

[6] K. Yesugade and R. Jadhav, “Implementation of Deep Learning Techniques for Deepfake Classification: A comparative study using ResNet-50 and VGG16,” in *2024 IEEE Pune Section International Conference Punecon 2024*, 2024. doi: 10.1109/PuneCon63413.2024.10894868.

[7] H. Gurram, K. Kethineni, D. H. Kommineni, and K. Soumya, “Gender Classification from Thermal Images using CNN,” in *2024 Asia Pacific Conference on Innovation in Technology Apcit 2024*, 2024. doi: 10.1109/APCIT62007.2024.10673553.

[8] H. Parikh, S. Patel, and V. Patel, “Transfer learning approach for patch level classification in remote sensing,” in *2019 IEEE 16th India Council International Conference Indicon 2019 Symposium Proceedings*, 2019. doi: 10.1109/INDICON47234.2019.9030283.

[9] D. Muduli, S. K. Sharma, D. Pradhan, M. Das, S. Mahapatra, and S. K. sahoo, *A Hybrid Deep Learning Approach for COVID-19 Detection: Feature Fusion and SVM Classification*, vol. 1153. 2025. doi: 10.1007/978-981-97-8093-8_8.

[10] J. Bozek, I. Kesedzic, L. Novosel, and T. Bozek, “Classification and feature analysis of the Human Connectome Project dataset for differentiating between males and females,” *Automatika*, vol. 62, no. 1, pp. 109–117, 2021, doi: 10.1080/00051144.2021.1885890.

[11] Z. S. A. Ghrban and N. K. EL Abbadi, “Gender Classification from Face and Eyes Images Using Deep Learning Algorithm,” *Journal of Computer Science*, vol. 19, no. 3, pp. 345–362, 2023, doi: 10.3844/JCSSP.2023.345.362.

[12] F. waris, F. Da, and S. Liu, “Stacked ensemble learning for facial gender classification using deep learning based features extraction,” *Cluster Comput.*, vol. 27, no. 8, pp. 11491–11513, 2024, doi: 10.1007/s10586-024-04340-7.

Informatics, Electrical and Electronics Engineering (Infotron)
Volume 6, Number 1, (2026) pp. 1-13
ISSN 2798-0197

[13] M. L. Sai Gayathri, N. Naga Kalyan, P. Vamsidhar, S. Lalitha, A. N. Sree Vaishnavi Neha, and M. Srinivas, "Real-Time Gender Classification from Facial Images Using VGG16 and GoogLeNet Architectures," in *2025 International Conference on Artificial Intelligence and Data Engineering Aide 2025 Proceedings*, 2025, pp. 158–164. doi: 10.1109/AIDE64228.2025.10987491.

[14] G. S. R. Tholeti, D. Ghanta, N. V. S. G. S. S. Chilukuri, and S. Bano, *Vocal Source Builds Divergence in Gender Recognition*, vol. 783. 2022. doi: 10.1007/978-981-16-3690-5_16.

[15] H. T. A. Ali and D. A. Nasrawi, "Extracting Gender Textual Nuances Using Text Similarity for Gender Classification Improvement," in *Icsintesa 2024 2024 4th International Conference of Science and Information Technology in Smart Administration the Collaboration of Smart Technology and Good Governance for Sustainable Development Goals*, 2024, pp. 72–76. doi: 10.1109/ICSINTESA62455.2024.10748083.

[16] P. Shourie, V. Anand, R. Chauhan, G. Verma, and S. Gupta, "A Deep Dive into Gender Classification Using Inception V3: Performance and Insights," in *2023 Global Conference on Information Technologies and Communications (GCITC)*, IEEE, Dec. 2023, pp. 1–6. doi: 10.1109/GCITC60406.2023.10426245.

[17] M. Kapse, N. Elangovan, M. Lalkiya, and A. Deshpande, *A Comparative Analysis of Machine Learning Algorithms for Image Classification: Evaluating Performance*. 2024. doi: 10.1007/978-981-97-2902-9_3.

[18] M. Rathi, A. A. Shah, S. Ali, E. N. Baro, A. W. Memon, and A. Inam, "Binary inception V3 deep learning based image classifier for the detection of breast cancer," in *Aip Conference Proceedings*, 2024. doi: 10.1063/5.0214634.

[19] Z. Jiang, "Face gender classification based on convolutional neural networks," in *Proceedings 2020 International Conference on Computer Information and Big Data Applications Cibda 2020*, 2020, pp. 120–123. doi: 10.1109/CIBDA50819.2020.00035.

[20] R. K. Yadav, R. Moriwal, and A. Sakhare, "Analysis of Computational Model for Detection and Recognition of Human Activity Using Deep Learning," in *2024 IEEE International Students Conference on Electrical Electronics and Computer Science Sceecs 2024*, 2024. doi: 10.1109/SCEECS61402.2024.10482159.

[21] A. S. Girsang, A. D. Saputra, and V. Yanrie, "Performance Comparison between VGG16 and Inception V3 for Organic Waste and Recyclable Waste Classification," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 11, no. 2, pp. 557–563, 2023.

[22] T. Reznychenko, E. Uglickich, and I. Nagy, "Accuracy Comparison of Logistic Regression, Random Forest, and Neural Networks Applied to Real MaaS Data," in *2024 Smart Cities Symposium Prague Scsp 2024 Proceedings*, 2024. doi: 10.1109/SCSP61506.2024.10552715.

[23] G. S. R. Tholeti, D. Ghanta, N. V. S. G. S. S. Chilukuri, and S. Bano, *Vocal Source Builds Divergence in Gender Recognition*, vol. 783. 2022. doi: 10.1007/978-981-16-3690-5_16.

[24] U. R. Vagalla and J. J. T. Selvaraj, "Customer segmentation of car retails using regularized logistic regression compared with K means clustering," in *Aip Conference Proceedings*, 2023. doi: 10.1063/5.0118059.

[25] S. Dong, "Research on the Enhancement and Improvement of Logistic Regression Models," in *2025 6th International Conference on Electrical Electronic Information and Communication Engineering Eeice 2025*, 2025, pp. 492–496. doi: 10.1109/EEICE65049.2025.11033904.

[26] E. Yücesoy, "Gender Recognition from Speech Signal Using CNN, KNN, SVM and RF," in *Procedia Computer Science*, 2024, pp. 2251–2257. doi: 10.1016/j.procs.2024.04.213.

[27] H.-S. Gao, A.-L. Guo, X.-D. Yu, and C.-C. Li, "RBF-SVM and its application on network security risk evaluation," in *2008 International Conference on Wireless Communications Networking and Mobile Computing Wicom 2008*, 2008. doi: 10.1109/WiCom.2008.1110.

[28] N. A. Sahadun, H. Haron, and M. R. A. Kadir, "Classification of gender using neural network and support vector machine from bone morphology properties of monkey," in *2013 IEEE Conference on Open Systems Icos 2013*, 2013, pp. 215–220. doi: 10.1109/ICOS.2013.6735077.

DOI: 10.33474/infotron.v6i1.24670

Informatics, Electrical and Electronics Engineering (Infotron)
Volume 6, Number 1, (2026) pp. 1-13
ISSN 2798-0197

[29] A. Olugboja, *Optimizing Predictive Performance: A Comparative Study of Machine Learning Algorithms for Classification Tasks*, vol. 2501 CCIS. 2025. doi: 10.1007/978-3-031-90341-0_11.

[30] M. Sathish Kumar, T. Sagithya, S. K. Arthi, and K. Subramaniam, "Empirical Evaluation of Activation Functions and Solvers on Neural Network Performance," in *Proceedings of the 9th International Conference on Communication and Electronics Systems Icces 2024*, 2024, pp. 1942–1946. doi: 10.1109/ICCES63552.2024.10859295.

[31] A. Mandal, N. Hasan, A. Jain, J. R. Mohanty, and V. Sharma, "Computational Relevance of Model Pruning and Quantization for Low-Powered AI," in *2025 2nd International Conference on Multidisciplinary Research and Innovations in Engineering Mrie 2025*, 2025, pp. 6–10. doi: 10.1109/MRIE66930.2025.11156676.

[32] X. Zhao *et al.*, "U-Net-based VGG19 model for improved facial expression recognition丨基于U-Net架构改进VGG19模型的人脸表情识别方法," *Gongcheng Kexue Xuebao Chinese Journal of Engineering*, vol. 47, no. 6, pp. 1272–1284, 2025, doi: 10.13374/j.issn2095-9389.2024.07.24.002.

[33] A. Bakhtiyar, M. A. Ansari, A. Mewada, and D. K. Singh, "From Pixels to People: Deep Learning Breakthroughs in Human Detection," in *Proceedings 2024 IEEE 16th International Conference on Communication Systems and Network Technologies Cicn 2024*, 2024, pp. 326–331. doi: 10.1109/CICN63059.2024.10847522.

[34] R. Wang, "Comparative Analysis of Deep Learning Models in Animal Species Recognition: A Case Study on the AFHQ Dataset," in *Aip Conference Proceedings*, 2024. doi: 10.1063/5.0223635.

[35] A. A. A. Aboluhom and I. Kandilli, "Face recognition using deep learning on Raspberry Pi," *Computer Journal*, vol. 67, no. 10, pp. 3020–3030, 2024, doi: 10.1093/comjnl/bxae066.

[36] Z. Cui, M. Zhang, and Y. Chen, "Deep embedding logistic regression," in *Proceedings 9th IEEE International Conference on Big Knowledge Icbk 2018*, 2018, pp. 176–183. doi: 10.1109/ICBK.2018.00031.

[37] S. Rajvanshi, G. Kaur, A. Dhatwalia, Arunima, A. Singla, and A. Bhasin, *Research on Problems and Solutions of Overfitting in Machine Learning*, vol. 1191 LNEE. 2024. doi: 10.1007/978-981-97-2508-3_47.

[38] J. P. Gygi, S. H. Kleinstein, and L. Guan, "Predictive overfitting in immunological applications: Pitfalls and solutions," *Hum. Vaccin. Immunother.*, vol. 19, no. 2, 2023, doi: 10.1080/21645515.2023.2251830.

[39] X. Li *et al.*, "From Embeddings to Accuracy: Comparing Foundation Models for Radiographic Classification," *Journal of Imaging Informatics in Medicine*, 2025, doi: 10.1007/s10278-025-01747-5.

DOI: 10.33474/infotron.v6i1.24670