Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 9, No. 2, April, 2025

Vision Transformer Embedded Feature Fusion Model with Pre-Trained Transformers for Keratoconus Disease Classification

Md Fatin Ishrak [1], Md Maruf Rahman [2], Md Imran Kabir Joy [3], Anna Tamuly [4], Salma Akter [5], Dewan M. Tanim [6], Shahajada Jawar [6], Nayeem Ahmed [4], Md Sadekur Rahman [7*]

[1] *Department of Electrical and Computer Engineering, University of Memphis, Memphis, United States.*

[2] *Department of Marketing & Business Analytics, Texas A&M University- Commerce, Texas, United States.*

[3] *MSA in Engineering Management, Central Michigan University, Michigan, United States.*

[4] *Department of Computer Science, University of Memphis, Memphis, United States.*

[5] *Department of Public Administration, Gannon University, Pennsylvania, United States.*

[6] *Department of Computer and Information Science, Gannon University, Pennsylvania, United States.*

[7] *Department of Computer Science and Engineering, Daffodil International University, Birulia, Bangladesh.*

Abstract

Keratoconus is a progressive eye disorder that, if undetected, can lead to severe visual impairment or blindness, necessitating early and accurate diagnosis. The primary objective of this research is to develop a feature fusion hybrid deep learning framework that integrates pretrained Convolutional Neural Networks (CNNs) with Vision Transformers (ViTs) for the automated classification of keratoconus into three distinct categories: Keratoconus, Normal, and Suspect. The dataset employed in this study is sourced from a widely recognized and publicly available online repository. Prior to model development, comprehensive preprocessing techniques were applied, including the removal of low-quality samples, image resizing, rescaling, and data augmentation. The dataset was subsequently partitioned into training, testing, and validation subsets to facilitate robust model training and performance evaluation. Eight state-of-the-art CNN architectures, including DenseNet121, EfficientNetB0, InceptionResNetV2, InceptionV3, MobileNetV2, ResNet50, VGG16, and VGG19, were utilized for feature extraction, while the ViT served as the classification head, leveraging its global attention mechanism for enhanced contextual learning, achieving near-perfect accuracy (e.g., DenseNet121+ViT: 99.28%). This study underscores the potential of hybrid CNN-ViT architectures to revolutionize keratoconus diagnosis, offering scalable, accurate, and efficient solutions to overcome limitations of traditional diagnostic methods while paving the way for broader applications in medical imaging.

Keywords:

Feature Fusion Model;
Keratoconus;
Vision Transformer; DenseNet121;
EfficientNetB0;
InceptionResNetV2; InceptionV3;
MobileNetV2; ResNet50; VGG16; VGG19.

Article History:

Received:	17	January	2025
Revised:	19	March	2025
Accepted:	26	March	2025
Published:	01	April	2025

1- Introduction

Keratoconus is a non-inflammatory, progressive corneal disorder characterized by corneal thinning and cone-shaped protrusion, which leads to visual impairment and, if untreated, can result in severe vision loss [1, 2]. Typically presented during adolescence, keratoconus causes irregular astigmatism, myopia, and distorted vision due to the irregular curvature of the cornea. This condition significantly affects patients' quality of life and imposes a considerable socio-economic

* **CONTACT**: sadekur.cse@daffodilvarsity.edu.bd

DOI: http://dx.doi.org/10.28991/ESJ-2025-09-02-027

burden [3]. In clinical practice, early and accurate detection of keratoconus is essential to halt disease progression and implement timely interventions, such as corneal cross-linking or corrective refractive procedures. However, the diagnosis of subclinical and early-stage keratoconus remains challenging due to its subtle clinical manifestations and the overlapping features it shares with normal corneas and other ecstatic disorders [4].

Traditionally, keratoconus detection has relied on corneal imaging techniques, including corneal topography and tomography, which provide valuable structural and morphological information. Devices such as Placido-based corneal topographers, Scheimpflug-based tomographers, and anterior segment optical coherence tomography (AS-OCT) have been instrumental in assessing key parameters like corneal curvature, thickness, elevation, and biomechanical properties [5]. These imaging modalities, combined with indices such as the Belin/Ambrosio Enhanced Ectasia Display (BAD-D) and Pentacam indices, offer essential diagnostic information. Nevertheless, accurately distinguishing between healthy eyes, keratoconus suspects, and confirmed keratoconus cases remains complex, especially in borderline and early-stage cases [3].

Despite advancements in diagnostic imaging, the manual interpretation of corneal maps remains subjective, time-consuming, and prone to inter-observer variability [4]. Moreover, while traditional machine learning (ML) models such as Support Vector Machines (SVM), Random Forest (RF), and Decision Trees (DT) have been explored for Keratoconus detection, these models largely depend on handcrafted features and domain expertise, limiting their generalizability and robustness. A significant gap persists in effectively leveraging the wealth of data produced by corneal imaging tools to capture subtle, non-linear patterns associated with early and subclinical Keratoconus [6].

Furthermore, existing models predominantly focus on detecting binary classifications (keratoconus vs. normal) without adequately addressing the need for accurate severity staging and progression prediction. The accurate staging of Keratoconus is crucial for developing personalized treatment plans and monitoring disease progression. This gap is further exacerbated by the limited adoption of advanced deep learning models like Transformers, which have shown promising performance in other domains but remain underexplored in ophthalmology, particularly in Keratoconus classification [7].

The advent of deep learning has revolutionized medical image analysis, offering transformative potential for automating disease detection and classification with unprecedented accuracy and efficiency [8]. Deep learning models, particularly Convolutional Neural Networks (CNNs), have become the cornerstone of modern automated medical imaging due to their ability to extract hierarchical features directly from raw data, bypassing the need for manual feature engineering [9]. CNN architectures such as DenseNet121 [10], EfficientNetB0 [11], ResNet50 [12], InceptionResNetV2 [13], InceptionV3 [14], MobileNetV2 [15], VGG16, and VGG19 [16] have demonstrated exceptional performance across a range of image classification tasks. These models leverage pretraining on large-scale datasets like ImageNet [17], which contains millions of labeled images, to initialize their weights before fine-tuning on specific medical datasets. This transfer learning approach has proven particularly valuable in medical imaging, where annotated datasets are often small due to the labor-intensive nature of expert labeling [18]. For keratoconus detection, CNNs can analyze corneal topography maps or optical coherence tomography (OCT) images to identify patterns indicative of the disease, such as localized steepening or thinning, with a level of precision that rivals or exceeds human experts [9].

Despite their success, CNNs have inherent limitations that can hinder their performance in complex medical imaging tasks like keratoconus classification. CNNs rely on convolutional layers with local receptive fields, which excel at capturing spatial hierarchies and local patterns—such as edges, textures, or small-scale irregularities in corneal images [8]. However, this localized focus restricts their ability to model long-range dependencies and global contextual relationships within an image [8]. In the context of keratoconus, understanding the broader spatial configuration of the cornea—such as the relationship between thinning in one region and steepening in another—is crucial for accurate diagnosis, especially in early or subclinical stages where local changes may be subtle [9]. The inability of CNNs to effectively capture these global interactions can lead to misclassification, particularly when distinguishing between normal corneas, subclinical keratoconus, and confirmed cases, which often share overlapping features [19]. Furthermore, the computational complexity of many CNN models, such as VGG16 or InceptionResNetV2, can make them impractical for deployment in resource-limited settings, where lightweight and efficient solutions are needed [11].

Additionally, ensemble learning approaches have gained traction in this field. Muhsin et al. introduced a two-stage ensemble learning framework that integrates multiple ML models, including RF, Gradient Boost (GB), DT, and SVM, achieving a validation accuracy of 99.41% in Keratoconus severity classification [6]. Similarly, transfer learning strategies leveraging pre-trained models like AlexNet, MobileNetV2, and ShuffleNet have demonstrated improved performance by fine-tuning on Keratoconus datasets [20].

However, despite these advancements, CNNs inherently possess a limitation: their local receptive fields restrict their ability to capture long-range dependencies and global contextual relationships within images [21]. This limitation is particularly critical in the case of Keratoconus, where the disease's subtle morphological patterns are spread across the corneal surface and require a holistic understanding for accurate classification.

The emergence of Vision Transformers (ViTs) has introduced a paradigm shift in image classification, addressing many of the shortcomings of CNNs by leveraging self-attention mechanisms [21]. Unlike CNNs, which process images through a series of convolutional filters, ViTs divide an image into fixed-size patches—typically 16x16 pixels—and treat them as a sequence of tokens, similar to how natural language processing models handle text [22]. These patches are then fed into a transformer encoder, where self-attention allows the model to weigh the importance of each patch relative to all others, capturing both local details and long-range dependencies simultaneously [8]. This global perspective makes ViTs particularly well-suited for medical imaging tasks like keratoconus detection, where the interplay between distant regions of the cornea can provide critical diagnostic clues [23]. For example, a ViT could relate a thinning patch in the central cornea to an elevation patch on the periphery, offering a more holistic understanding of the disease's topographic signature than a CNN's localized analysis. However, ViTs require large-scale datasets for effective training, posing a challenge in specialized domains like ophthalmology, where annotated data is limited. To overcome this, researchers have adopted pre-trained Transformer models and fine-tuned them on domain-specific datasets, leveraging transfer learning to retain performance while addressing data scarcity [7].

Moreover, feature fusion techniques that combine features extracted from multiple models have proven effective in enhancing classification performance. By integrating the complementary strengths of CNNs and Transformers, hybrid models can leverage local feature extraction and global contextual understanding, providing a more comprehensive analysis of corneal imaging data [2, 7]. Given the demonstrated success of ensemble learning, transfer learning, and deep feature fusion in KC detection, the motivation arises to explore a hybrid model that embeds Vision Transformers within a feature fusion framework, thereby enhancing keratoconus classification performance and addressing existing gaps.

In this study, we propose a novel Vision Transformer Embedded Feature Fusion Model for Keratoconus Disease Classification. Our approach integrates pre-trained Transformer architectures with feature fusion techniques to classify KC severity stages accurately. Unlike conventional CNN-based methods, our model captures both local and global patterns across corneal maps, enabling robust detection of early, subclinical, and advanced KC cases.

Specifically, our contributions include:

- Novel Hybrid Architecture: The proposed model integrates the local feature extraction strengths of pretrained CNNs with the global contextual modeling capabilities of Vision Transformers.
- Global Contextual Fusion: Vision Transformers enhance the model's ability to capture long-range dependencies and contextual patterns, addressing CNNs' limitations.
- Comprehensive Evaluation: Rigorous performance evaluation across multiple metrics provides a holistic understanding of the model's classification capability and robustness.

The remainder of this paper is organized as follows: Section 2 reviews the related work on Keratoconus detection using machine learning and deep learning techniques. Section 3 details the methodology, including dataset description, model architecture, and feature fusion strategy, the experimental setup and evaluation metrics. Section 4 discusses the findings and analysis of the findings. Finally, Section 5 concludes the paper and outlines future research directions for enhancing keratoconus diagnosis using advanced deep learning models.

2- Literature Review

Keratoconus categorization, as a degenerative ocular condition that affects the configuration and optical performance of the cornea, has been an increasingly important concern in medical image analysis. Recent advances in artificial intelligence, especially in deep learning techniques, have provided very strong tools for automating this classification task. This literature review considers the current methodologies, challenges, and new strategies related to keratoconus classification.

2-1- Deep Learning in Medical Image Classification

Deep learning has revolutionized the field of medical image analysis and thus allows the automation of complex tasks related to disease diagnosis, classification, and monitoring. Unlike traditional machine learning approaches, which rely on handcrafted features, deep learning models-especially CNNs-automatically learn hierarchical patterns from raw data. This brings about proficiency that makes them particularly well-suited for handling the complex structures and patterns in medical imaging, including those from X-rays, MRIs, CT scans, and corneal topography [2]. Such developments have greatly improved the accuracy and effectiveness of medical diagnoses, thus circumventing the limitations of traditional methods.

The development of deep learning in medical imaging was able to be done with complex CNN architectures and large-scale datasets. Early CNNs, such as LeNet by LeCun et al. [24] formed the bedrock for deep learning applications. Further development in advanced architectures like VGG by Simonyan & Zisserman [16], ResNet by He et al. [12], and DenseNet [10] constituted a great stride in the evolution of CNN architecture. These architectures have leveraged transfer learning, where models pre-trained on large datasets like ImageNet are fine-tuned on domain-specific medical datasets. This would, therefore, address the problem of limited labeled medical data and ensure high generalizability across diverse datasets [25].

One of the reasons deep learning has come to the fore in medical imaging is its ability to do automatic feature extraction. In conventional machine learning models, domain experts need to design and extract features manually, which can be extremely time-consuming and prone to errors. On the other hand, CNNs extract hierarchical features directly from raw images. For example, the early layers of a CNN detect basic features, including edges and textures, while the higher-level layers take part in the recognition of more complex semantic features, such as shapes and structures relevant to specific medical conditions [26]. This hierarchical feature extraction capability makes CNNs especially good at detecting and classifying abnormalities in medical imaging, such as lesions, tumors, or corneal deformities.

The pretrained CNN models are getting popular due to their robustness and flexibility in medical image classification. For example, VGG networks have a very simple yet deep architecture, making them quite effective in the analysis of high-resolution medical images [16]. Overcoming the vanishing gradient problem, ResNet with its residual connections enables training arbitrarily deeper networks and achieves state-of-the-art performance in a wide range of abnormalities, including lung nodule detection and diabetic retinopathy diagnosis [12]. The idea of feature re-use and reduction of parameter redundancy through dense layer connections is further extended by DenseNet, which helps in small medical datasets [10]. Conversely, EfficientNet unifies the model dimensions of depth, width, and resolution to achieve state-of-the-art accuracy using less computational resources, which makes it especially suitable for resource-constrained environments.

Convolutional Neural Networks have achieved significant success in medical imaging, yet they face key limitations. Their focus on localized features often overlooks important global contextual information, such as the overall curvature of the cornea crucial for keratoconus classification. Moreover, CNNs require large amounts of labelled data, making expert annotation resource intensive. Their "black-box" nature also limits interpretability, posing challenges for clinical trust and adoption. While CNNs excel in tasks like disease classification and tumour segmentation, the emergence of Vision Transformers addresses these issues by using self-attention mechanisms to capture both local and global dependencies, improving interpretability and reducing data demands [21].

2-2- Vision Transformers in Medical Imaging

Recently, Vision Transformers have emerged as a game-changing architecture in medical imaging, allowing a complete paradigm shift in image analysis by robustly leveraging self-attention mechanisms in capturing long-range dependencies and global contextual information. This capability mitigates some of the traditional CNN limitations in modelling complex structures inherent in medical images.

A review by Xu et al. has identified with precision that ViTs have been adopted for several medical image analysis tasks such as classification, segmentation, detection, registration, synthesis, and clinical report generation. It underlines the superior ability of ViTs to understand global contextual interrelations-crucial to make precise interpretations of medical images. The authors also discuss challenges in the use of ViTs, including large datasets and computational resources required, and future research directions to make them more applicable in medical imaging [27]. Islam et al. present an extended review on transformer applications in medical imaging and perform a comparison with CNNs in various tasks. According to the review, transformers bear huge promise for outperforming CNNs in specific scenarios, especially for the capture of complex anatomical structures and long-range dependencies. They further discuss challenges of using transformers in medical imaging, including limited large and annotated datasets, and high computational requirements, enumerating the strategies to overcome such barriers [28].

Al-hammuri et al. give an overview of the basic principles of the ViT architecture and its applications for image segmentation, classification, detection, prediction, reconstruction, synthesis, and telehealth in digital health. According to this article, the advantages of using ViTs for the processing of medical images are that these models can process long-range dependencies and spatial correlations-features very instrumental in making the right diagnosis and subsequent treatment plan. The authors further discuss the limitations of the ViTs, needing large datasets and computational resources, and propose future research directions to improve their applicability in digital health [29]. One such research investigation, by Islam et al. [28], relates to the use of ViTs in lung cancer imaging in an attempt to outline their capability to improve diagnostic and prognostic outcomes. In fact, excellent performance in capturing intricate patterns or long-range dependencies in medical images, quintessential for the proper detection and classification of lung cancers, justifies the use of ViTs. Moreover, it brings to light the challenges that stand in the way of employing ViTs for clinical practice, such as high computational demands or the need for large and annotated datasets and proposes several ways to surpass these obstacles. Chen et al. [2] proposed 3D TransUNet-a model that includes ViTs integrated into the U-Net for medical image segmentation. This work showed how the integration of ViTs in a strong model further enhances global context representation, hence yielding better segmentation performance for complex tasks. The authors also discussed some challenges when integrating ViTs into existing architectures, such as higher computational complexity, and gave some possible solutions to these issues.

2-3- Hybrid Deep Learning Models in Medical Imaging

Hybrid deep learning frameworks have been gaining much attention in the area of medical imaging, as they combine the benefits of traditional machine learning approaches with those of deep learning architectures. The goal of such models is to overcome the limitations posed by isolated techniques, combining handcrafted feature extraction with layered feature learning capabilities characteristic of deep learning. This has been shown to be very effective in medical imaging, particularly in applications where datasets are typically small, and the image patterns complex and variable [30].

Conventional traditional machine learning-based approaches to medical imaging included methods using SVMs and random forests. Conventionally, such schemes have been reliant on handcrafted features-developed either manually or using well-identified algorithms-but are inescapably far too time-consuming and cumbersome for hand Derivation, not to mention an inability even to characterize relatively subtle or sophisticated pattern variants found in many of these medical images. While deep learning models, especially CNNs, are amazingly capable of automatic feature learning from scratch in raw images, hybrid models integrate both approaches, combining the strengths of both high accuracies provided by handcrafted features with the depth and adaptability of features which could be learned by deep neural networks [31].

Hybrid models have found a wide range of applications in various tasks related to medical imaging. For example, Srinivasan et al. [32] suggested a hybrid deep CNN model for the classification of brain tumours. Traditional image preprocessing was combined with deep learning in this model. This model had a detection accuracy of 99.53%, hence showing its promise for the early diagnosis of brain malignancies. Similarly, Tamilselvi et al. [33] proposed a hybrid model that combined Swin Transformer and ResNet50 architectures for the automated diagnosis of eye diseases. The combination enhanced the performance of classification, thus demonstrating the advantages of multiple modelling techniques being combined. In medical image segmentation, Azad et al. [34] reviewed hybrid approaches that combined traditional segmentation techniques with deep learning. These models have performed excellently in segmenting complicated anatomical structures, thus proving their adaptability to different tasks.

Hybrid models offer several significant advantages in medical image analysis. By combining the strengths of traditional machine learning and deep learning, they achieve higher diagnostic accuracy—leveraging robust handcrafted features from traditional methods alongside the hierarchical feature extraction capabilities of deep learning, which captures both local and global contexts. Additionally, hybrid models enhance interpretability, as traditional algorithms provide a clearer rationale for decisions, partially mitigating the "black box" issue of deep learning. They are also more data-efficient, performing well even with limited annotated datasets by enriching the feature set. However, hybrid models present challenges. Designing and implementing them requires expertise in both machine learning paradigms, increasing complexity. Furthermore, these models can be computationally intensive, potentially limiting scalability in real-world applications. Future research should focus on simplifying the integration of traditional and deep learning approaches, optimizing computational efficiency, and enhancing generalizability across diverse tasks in medical imaging.

2-4- Feature Fusion for Medical Image Classification

Feature fusion plays an important role in medical imaging classification, which allows the integration of complementary information for better diagnostics with high accuracy and robustness. Such techniques also avoid the intrinsic limitations arising in single-modality analyses, in that one modality often lacks the whole range of relevant details related to both anatomy and functions. Feature fusion refers to the combination of features from different models or modalities, offering an inclusive data representation for better decision-making. As an example, the combination of structural information from MRI with functional data from PET creates a holistic view of patient conditions, hence improved classification and detection results [35].

Features can also be fused at higher levels in the classification pipeline. Early fusion is a technique of first combining raw data from multiple modalities before the extraction of features therefrom. Methods such as this do allow the capture of low-level correlation among multiple data sources but probably introduce noise and redundancy that may confound learning too. In some applications, multi-modal fusion involves independent feature extraction in each modality, then further combines them. This effectively fuses complementary information while preserving the integrity of individual features. For example, fusion of texture features from MRI and metabolic activity features from PET resulted in improved accuracy of tumour classification as shown by Xie et al. [10]. Late fusion or decision-level fusion combines outputs from individual models or classifiers, using techniques like majority voting or weighted averaging. This will be helpful if the performance of individual classifiers is good, and they cannot take advantage of the complementary aspects from different sets of features [36].

Feature fusion finds an extensive range of applications in medical imaging, with high impacts. It has been successfully applied to the detection and classification of brain tumours by providing both anatomical and functional insights through fusion of MRI-PET data among others [37]. Hybrid models that integrate CNNs with ViTs show great promise for classifying retinal diseases. These architectures are based on the complementary properties of CNNs for the capture of local features and ViTs for global dependencies modelling, such as by Matsoukas et al. [38]. Multi-scale analysis enables the capture of both fine and coarse details by feature fusion from different layers of CNNs, thereby enhancing tasks such as histopathological image classification by Yamanakkanavar et al. [39].

However, besides the advantages, there are also several challenges in feature fusion. Data heterogeneity can make the process of fusion difficult, which involves resolution, contrast, and noise across modalities. The robustness of some preprocessing techniques may be necessary to alleviate these issues. Feature fusion challenges the development of algorithms and hardware solutions that are able to integrate multiple feature sets with high computational complexity. Moreover, interpretability remains a significant concern, as understanding the contribution of fused features to the final decision is critical in clinical settings. Explainable AI techniques can help address this issue, providing insights into the decision-making process of fused models [40, 41].

2-5- Models for Keratoconus Classification

Lavric and Valentin proposed the KeratoDetect algorithm based on the Convolutional Neural Network (CNN) to classify keratoconus from corneal topography images. For this, 1,500 images are considered in the dataset for healthy eyes and 1,500 images for affected eyes by keratoconus. The model was trained using 1,350 images, its validation done on 150 images, and testing done on 200 images. The accuracy achieved on the test dataset was 99.33%. It was a binary classification of either healthy eyes or those with keratoconus. The high accuracy ensures the potential for deep learning algorithms like CNNs to improve diagnostic precision in the diagnosis of keratoconus [42].

Al-Timemy et al. have used the Xception and InceptionResNetV2 models for keratoconus classification, based on the Egyptian and Iraqi datasets consisting of a total number of 4,752 corneal images. The dataset included three classes: normal, suspect keratoconus, and keratoconus. Models were combined in a feature fusion approach and integrated with classifiers like decision trees (DT) and support vector machines (SVM). The accuracy of classification was 97-100% for the main dataset and 88-92% for the independent test dataset of the fusion model. The approach proposed will thus adhere to high efficacy in clinical and subclinical detection of keratoconus due to the inclusion of strong features [7].

Chen et al. classified keratoconus using a CNN on 1,926 corneal tomography scans divided into five classes: healthy and Amsler-Krumeich stages 1 to 4. Images in the training and testing datasets came from centers in the UK, New Zealand, and Iran, while independent validation was performed on the Iranian dataset. The CNN achieved an accuracy of 99.07% for distinguishing healthy eyes from keratoconus and 93.12% for grading disease severity. This model integrated color-coded axial, pachymetry, and elevation maps, showing the potential of CNN for reliable classification of keratoconus [2].

The unsupervised machine learning algorithm in Yousefi et al. classified the severity of keratoconus using 3,156 corneal OCT images. In their study, the data were divided into four clusters: normal eyes, suspected keratoconus, mild keratoconus, and advanced keratoconus based on 420 corneal topography, elevation, and pachymetry parameters. Dimensionality reduction by PCA was used, followed by t-SNE mapping and density-based clustering. This approach yielded 97.7% sensitivity and 94.1% specificity in binary classification. This post shows the potential of unsupervised learning in the effective staging of keratoconus without any need for labelled datasets [43].

Kamiya et al. proposed a CNN model based on VGG-16 using Placido disk-based corneal topography images to classify keratoconus, both at disease and between four grades using the Amsler-Krumeich classification. They considered 349 images: 179 representing keratoconus and 170 normals. The proposed model achieved an accuracy of 96.6% for keratoconus-normal discrimination and 78.5% in disease staging. Classification addressed five classes: normal and Grades 1 to 4. This study underlines the potential of CNNs in keratoconus detection and staging using widely available topography data [5].

Kuo et al. conducted a study to classify keratoconus using 354 corneal topography images with three deep learning models: VGG16, InceptionV3, and ResNet152. The dataset consisted of three classes: normal, keratoconus, and subclinical keratoconus. Among the three models, ResNet152 had the best performance, with an accuracy of 95.8%, sensitivity of 94.4%, and specificity of 97.2%. These visualization techniques-Grad-CAM, for example-offered insights into the diagnostic features employed by the models and therefore enhanced their clinical interpretability. This study illustrated the performance of convolutional neural networks in keratoconus screening and its grading [4].

Al-Timemy et al. proposed a hybrid deep learning model, comprising the architecture of EfficientNet-B0 combined with support vector machines for the classification of keratoconus. The model utilized 4,844 images derived from corneal maps across seven features, including anterior curvature and posterior elevation, to address three classes: normal, suspected keratoconus, and keratoconus. It achieved an accuracy of 98.8% for binary classification (normal versus keratoconus) and 81.5% for three-class classification in the development dataset, with independent validation of 92% and 68.7%, respectively. Thus, the paper presented here places a greater emphasis on hybrid architectures for the effective diagnosis of keratoconus [20].

The proposed study by Al-Timemy et al. [44] presented an Ensemble Deep Transfer Learning (EDTL) method that included some pretrained networks, such as SqueezeNet, AlexNet, ShuffleNet, and MobileNetV2, for detecting keratoconus using 2,136 corneal topographic maps. These were derived from 534 cases divided into two classes: keratoconus and normal. A probability fusion step was employed, combining results from four deep networks and a Pentacam Indices classifier. The EDTL has achieved 98.3% accuracy on AlexNet architecture when product fusion was applied for classification, hence showing performance with regard to multi-level feature combination and probability fusion in the robust identification of keratoconus.

Wan et al. explore machine learning-based prediction of keratoconus progression post-accelerated corneal collagen cross-linking (A-CXL), analyzing 95 eyes with a 3-22 month follow-up. Using LASSO, XGBoost, and random forest, they identify maximal keratometry (Kmax) and index of surface variance (ISV) as key prognostic factors, achieving a 98% accuracy with an XGBoost model. A nomogram enhances prediction over single parameters. However, the study's small, single-center sample and short follow-up limit generalizability. Future multi-center research is needed for broader applicability [45].

Hashim & Mazinani introduce a CNN-based model for keratoconus detection using corneal topography data from the Jenna Ophthalmic Center, achieving an exceptional accuracy of 99%. The study leverages a novel dataset and employs PCA for feature extraction, enhancing the model's performance over traditional ML methods like AdaBoost (70.9% accuracy). While the three-stage process—pre-processing, feature extraction, and classification—demonstrates robustness, the small dataset (528 images) and single-center focus limit generalizability. Scalability and real-world clinical integration remain unaddressed. Nonetheless, the high accuracy marks a significant advancement in early keratoconus detection [46].

Ismael contributed to keratoconus detection with Transformer Technology and Multi-Source Integration, with a novel transformer-based model for detecting keratoconus progression, integrating corneal topography, aberrometry, pachymetry, and biomechanical data. The proposed model achieves an impressive accuracy of 98.48% on the "Keratoconus Detection" dataset, surpassing SVM (93.45%), Random Forests (93.39%), and CNN (94.37%) methods. While the study excels in multimodal data fusion and early detection, its reliance on resource-intensive transformers may limit clinical scalability. Future work should address generalizability across diverse populations and real-time applicability [47].

Askarian et al. propose a novel smartphone-based method for detecting keratoconus (KC) using Placido disc projections and a Support Vector Machine (SVM) classifier, achieving a high accuracy of 97%. This standalone approach eliminates the need for costly equipment, enhancing accessibility in resource-limited settings. Strengths include robust preprocessing and adaptive contrast mechanisms, yielding 96.08% sensitivity and 97.96% specificity. However, the study's reliance on a small dataset of emulated eye models limits generalizability, necessitating real-patient validation. Despite this, the method offers a promising, cost-effective screening tool for early KC detection [48].

Hartmann et al. developed a neural network to predict keratoconus progression at the first visit, integrating Scheimpflug imaging and clinical data, achieving a highest accuracy of 0.83 using MobileNet and multilayer perceptron. The model excels with high specificity (0.95) but moderate sensitivity (0.53), suggesting reliable detection of nonprogressive cases yet missing some progressive ones. Limited by a small dataset (570 eyes) and retrospective design, generalizability is constrained. Ablation studies highlight age and posterior elevation as key predictors. Future multicenter studies could enhance robustness [49].

Yaraghi & Khatibi propose a novel keratoconus classification system integrating features from pretrained models (InceptionResNetV2, VGG16, EfficientNet-B0) with a vision transformer, achieving a highest accuracy of 96.20% on a public dataset. Evaluated on the Shahroud Cohort Eye dataset, it outperforms individual models (e.g., EfficientNet-B0: 85.72%) via feature fusion and transformer architecture. However, the small dataset (92 images) and reliance on corneal thickness maps limit generalizability. Data augmentation mitigates imbalance, but rotation exclusion may overlook critical variations. Future multi-map integration could enhance robustness [50].

The article by Shi et al. introduces FVAnet, a novel deep learning model for keratoconus detection using corneal topography maps, achieving a remarkable 99.7% training accuracy and 78.7% test accuracy. While its feature vector aggregation and pruning strategy enhance efficiency and generalization, the reliance on a limited dataset restricts broader applicability. The model's innovative angle-based classification outperforms traditional methods, yet its performance on diverse, larger datasets remains untested. Future research should address these limitations and explore severity staging [51].

2-6- Research Gap

Despite advancements in keratoconus classification using deep learning, significant gaps remain those are addressed in this research:

Enhancing Multimodal Data Integration

Although features extracted from different imaging modalities have shown the potential of combination, such as anterior elevation, posterior curvature, and pachymetry, the exploration toward advanced fusion strategies in order to maximize their complementary information remains limited. Therefore, our work intends to develop a more sophisticated feature fusion model that integrates more multimodal data-including additional corneal biomechanical parameters-and patient demographics in order to increase diagnostic accuracy and robustness.

Improving Performance in Multi-Class Classification

Most of the existing models give very high accuracy in binary classification, such as normal versus keratoconus, but have poor performance when dealing with multi-classification, like differentiating between normal, subclinical, progressive, and advanced keratoconus. Our research will be focused on the optimization of feature fusion techniques to handle multi-class challenges more effectively, ensuring high accuracy with reliable predictions across all stages of keratoconus.

3- Proposed Methodology

This research involves the development of a hybrid deep learning model for classifying keratoconus into three categories: Keratoconus, Suspected, and Normal. The methodology integrates feature extraction using pretrained convolutional neural network (CNN) models and classification using Vision Transformers (ViTs). Figure 1 presents the overall methodology of the research and in the subsequent sections a detailed explanation of the steps involved in the methodology is provided:

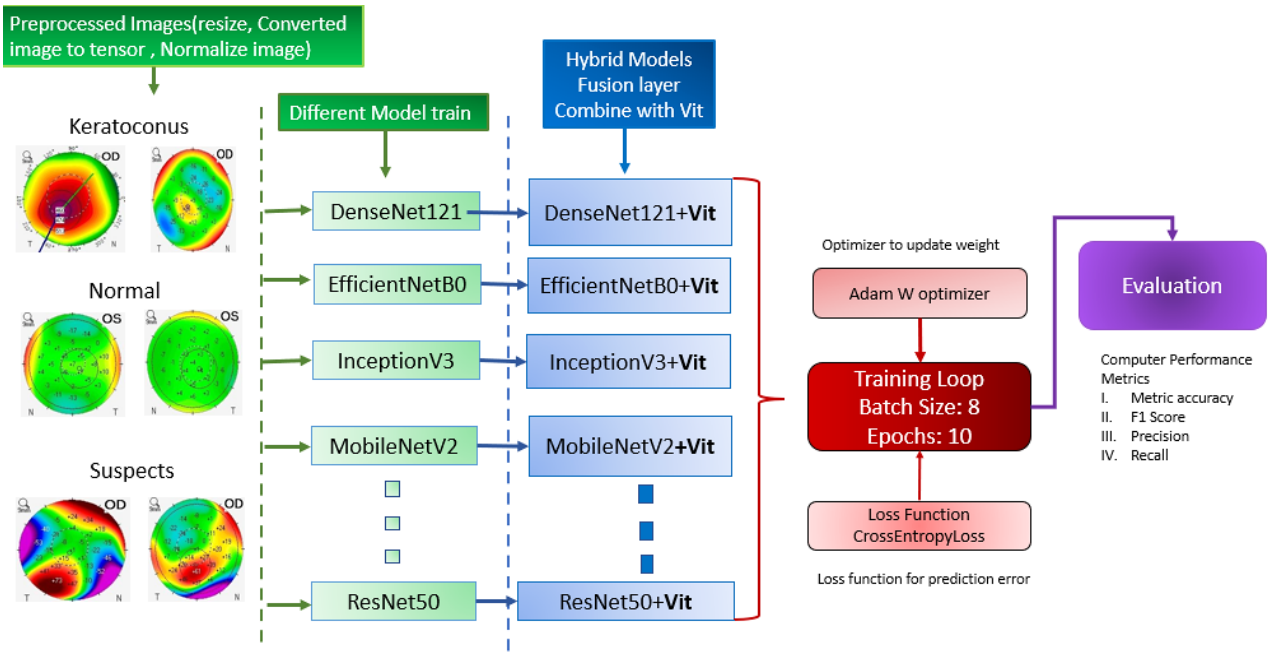
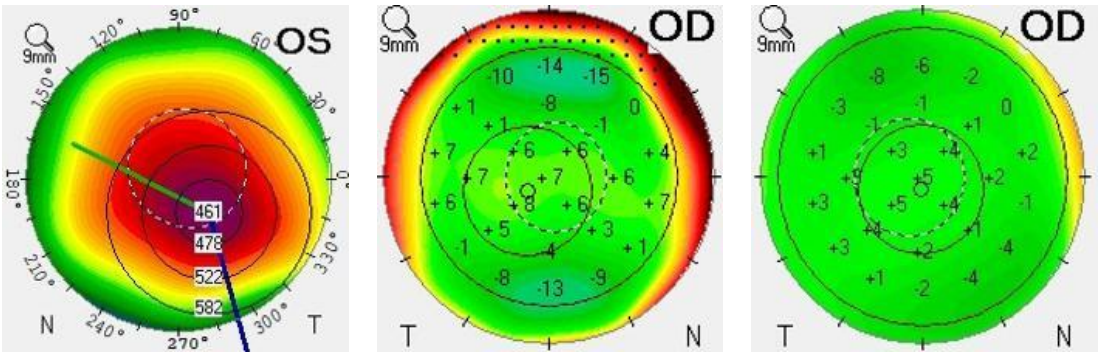


Figure 1. Proposed methodology of the work

3-1- Description of the Dataset

The secondary dataset was collected from online for keratoconus detection consists of 4,011 corneal topography images categorized into three distinct classes: Keratoconus, Normal, and Suspect. These images are essential for training and evaluating deep learning models that aim to classify and diagnose keratoconus, a progressive eye disorder. The dataset is well-structured but presents challenges that must be addressed to ensure optimal model performance. A few samples of the dataset are shown in Figure 2 and the distribution of the dataset is shown in Figure 3.



Keratoconus Suspected Normal

Figure 2. Sample dataset

This dataset contains two classes: Keratoconus and Normal, with 1,400 samples in each category and thus having equal proportions of 34.9% each. The Keratoconus class contains the corneal topography maps where features of keratoconus were definitely visible, like thinning of the cornea or cone-shaped protuberance of the cornea. On the other hand, the Normal class comprises pictures of a healthy cornea and, therefore, presents the ground for comparison. The Suspect class, consisting of 1,211 samples (30.2%), corresponds to cases in which there is some kind of abnormality that is not definitely keratoconus but may point toward the condition. This class complicates the classification problem since it falls between the normal and diseased states.

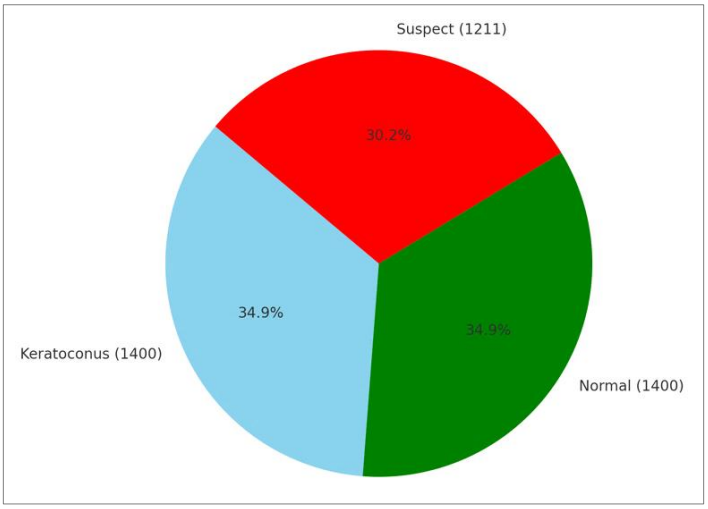


Figure 3. Distribution of the original dataset

Images in this dataset are, in fact, corneal topography maps that contain several features which relate to the curvature or thickness of the cornea, the key to the detection of abnormalities in keratoconus. Besides the subtleness innate between Normal and Suspect class discrimination, higher levels of advanced feature extraction along with sophistication required for high-performance classification in modeling are to be used accordingly.

Considering the dataset size is moderate, with 4,011 samples, data augmentation is necessary to enhance model robustness. The data augmentation strategy of rotation, flipping, scaling, and brightness will increase the dataset size, especially for the underrepresented class, Suspect, and make balanced learning possible. Hybrid approaches that leverage the strengths of pertained CNNs on feature extraction and those of ViTs for classification can also make more effective discriminations among classes by using both local and global features.

3-2- Data Preprocessing

Data preparation is necessary for deep learning model training. Model learning and generalization depend on data quality. The dataset was preprocessed by removing weak samples, reducing images, scaling pixel values, and adding data. Stages boost dataset quality and diversity, making the model more resilient and accurate.

3-2-1- Remove Bad Samples

Bad samples can damage model training. Photos may be blurry, mislabeled, corrupted, or other class-unsuitable artifacts. These samples must be eliminated to avoid the model from learning incorrect patterns or being confused by data noise. The collection was cleaned of blurry, poorly resolved, or mislabeled pictures. This ensures the dataset contains only high-quality samples, helping the model learn and reducing noisy data overfitting.

3-2-2- Resize

CNN models' input size was normalized by downsizing dataset photos. Model size is usually constant to interpret input images uniformly throughout network levels. Dataset photographs must be resized to 224x224 pixels if the model architecture supports them. Images are resized to preserve spatial structure and reduce computational and memory load. To avoid image distortion and losing important features, retain the aspect ratio when resizing.

3-2-3- Rescale

Image pixel values are rescaled to a given range, usually between 0 and 1 or -1 and 1. In this dataset, pixel values were rescaled from 0 to 255 for 8-bit images to 0 to 1. Normalizing input features ensures a uniform scale, speeding up training convergence. It also reduces the possibility of huge gradients or vanishing gradients from unnormalized pixel values. Rescaling is crucial for models pre-trained on big datasets like ImageNet, which require a certain input range.

3-2-4- Augmentation

Photo data augmentation expands and diversifies the training dataset. This dataset may have employed random rotations, flips, zooms, shifts, and brightness changes. These changes make the model more orientation, scale, and lighting invariant, improving its generalization to new data. Addition of samples for underrepresented classes reduces class imbalance, notably in medical imaging applications with few samples. Augmenting the training dataset reduces overfitting and improves model robustness. Each augmented image retains categorization but adds variants to reduce model input change. The outcome of the augmentation is shown in Figure 4.

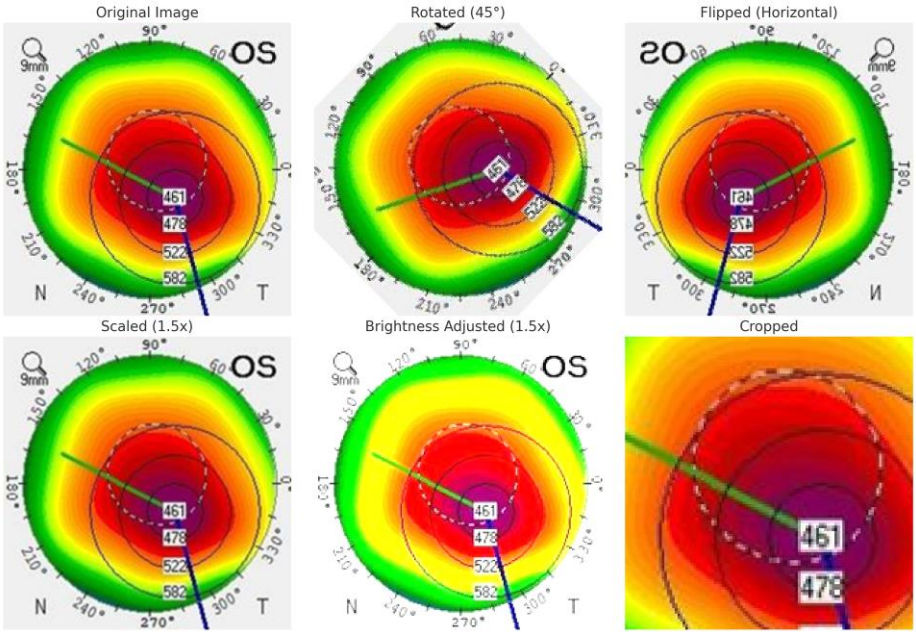


Figure 4. Sample outcome of the augmentation

3-2-5- Data Split

After preprocessing and augmentation, the keratoconus detection dataset has 16,016 samples divided into Normal, Suspect, and Keratoconus classes. To train and evaluate the model, the dataset was separated into training, validation, and testing sets. The training set has 9609 samples. Among them 3343 Keratoconus, 3360 Normal, and 2906 were Suspect samples. 3204 samples were taken as Validation set to tweak the model's hyperparameters and prevent overfitting. Samples include 1146 Keratoconus, 1094 Normal, and 964 Suspect. This set evaluates the model on unseen data during training. Finally, 3204 samples were taken for the test set. Its distribution matches the validation set: 1146 Keratoconus, 1094 Normal, and 964 Suspect. The test set evaluates model generalization to new data.

3-3- Feature Extraction and Model Building

Feature extraction and model building are the two main stages involved in the construction of a robust model for detecting keratoconus. The former relies on taking advantage of the pre-trained CNNs in extracting meaningful patterns from the corneal topography images, while the latter feeds these features into a hybrid deep learning model based on CNN and ViT for comprehensive classification.

3-3-1- Feature Extraction with Pretrained Models

Feature extraction in machine learning is finding and distilling meaningful patterns within raw data in order to best represent the true underlying structure of the information contained within. Eight different convolutional neural network models, pre-trained on large-scale datasets such as ImageNet, have been used in this work, thereby enabling rich, hierarchical image representations. This is the strongest point of each model, complementing one another in image feature extraction to assure completeness of represented data.

3-3-1-1- DenseNet121

DenseNet121 is an advanced deep learning network architecture designed for solving major issues in the training of deep networks, including vanishing gradient problem, overfitting, and redundancy in parameters. It was further developed by Huang et al. [8], having a unique connected structure, with every layer connected directly to all the previous layers. Further, it has grown very effective for feature extraction in medical image analysis and especially so suitable for such a complex task of keratoconus detection. Architecture of DenseNet121 is shown in Figure 5 [52].

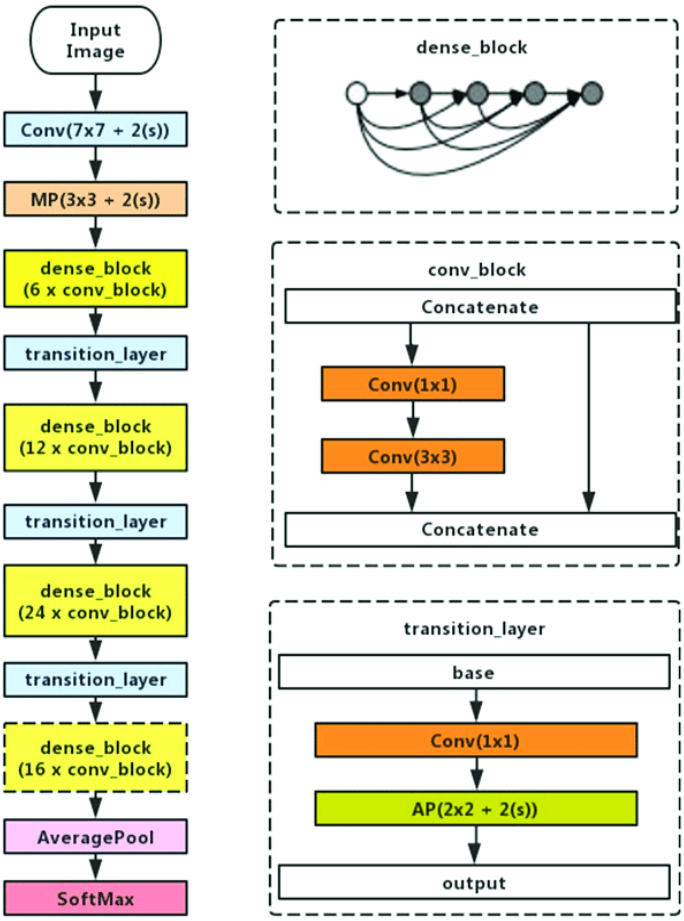


Figure 5. Architecture of DenseNet121

The main point about DenseNet121 is that its layers are densely connected; that is, each layer takes as input the feature maps of all the preceding layers and sends its feature maps to all subsequent layers. This grants great avarice for efficient feature re-usage since the model can just build upon previously learned features. With this mechanism, the model could learn various and rich representations of input data and enable it to learn subtle patterns in the corneal images.

Another important feature is the growth rate, k, which defines how many feature maps are added by each layer. For DenseNet121, the growth rate is k = 32, meaning that each layer adds 32 new feature maps. This controlled growth enables the model to maintain computational efficiency without sacrificing too much in feature quality. Also, the use of bottleneck layers with 1×11 convolutions before 3×33 convolutions reduce the dimensionality of the input feature maps and diminishes computation without losing too much accuracy.

Between these dense blocks, transition layers are utilized for subsampling in the feature maps. These transition layers consist of 1×11 convolutions followed by 2×22 average pooling; this reduces the spatial dimensions and overfitting. At the end of the network, a global average pooling layer aggregates the spatial information, while the fully connected layers are replaced. This architecture keeps the number of parameters low and makes the model robust and efficient.

The DenseNet121 network consists of 121 layers, comprising convolutional layers, bottleneck layers, dense blocks, and transition layers. There are four dense blocks separated by transition layers, and the layers in each dense block are densely connected in a feed-forward manner, meaning that information can propagate across the block. The growth rate of 32 controls and efficiently expands the feature maps.

It also has a global average pooling layer at the end, which aggregates spatial information from feature maps, and then uses a fully connected output layer for classification. The proposed architecture provides both high accuracy with computational efficiency.

3-3-1-2- EfficientNetB0

EfficientNetB0, proposed by Tan and Le, represents a state-of-the-art deep learning architecture with optimized accuracy and efficiency. It proposes a novel compound scaling method to systematically scale up the model depth, width, and resolution with a balanced trade-off between computational resources and performance. This novelty has attracted researchers' interest toward EfficientNetB0 for feature extraction in medical imaging, including keratoconus detection, where computational efficiency and accuracy are crucial [3].

The most salient feature of EfficientNetB0 is the mode of compound scaling. While traditional CNNs scale the depth-width-resolution independently of each other, EfficientNetB0 does so uniformly based on a compound scaling formula. This makes sure that computational resources are optimally allocated with improved accuracy and without additional complexity.

EfficientNetB0 also uses depthwise separable convolutions, factoring the convolution process into spatial filtering and channel-wise filtering. This greatly reduces the computational cost compared to a standard convolution. It also involves squeeze-and-excitation layers that carry out feature-map recalibration by emphasizing important channels. This mechanism helps the model focus on the most relevant features in the data.

The next in line would be the Swish activation instead of the traditional ReLU activation. Swish provides a much smoother gradient flow and allows the network to learn complicated patterns much better. Further, this is combined with the use of AutoML optimization, the architecture of EfficientNetB0 tries to reach a high level with relatively lesser computational overhead.

EfficientNetB0's architecture is modular and shown in Figure 6 [53]. A few key components of the model are described below:

- Input Layer: This layer resizes all images to a resolution of 224×224 pixels, ensuring compatibility with the model.
- Mobile Inverted Bottleneck Convolution (MBConv) Blocks: These blocks integrate depthwise separable convolutions and squeeze-and-excitation mechanisms to process the features efficiently.
- Global Average Pooling: This layer aggregates spatial information, reducing the dimensionality of feature maps while preserving essential details.
- Fully Connected Output Layer: The extracted features are processed into a compact representation, ready for further analysis in the hybrid model.

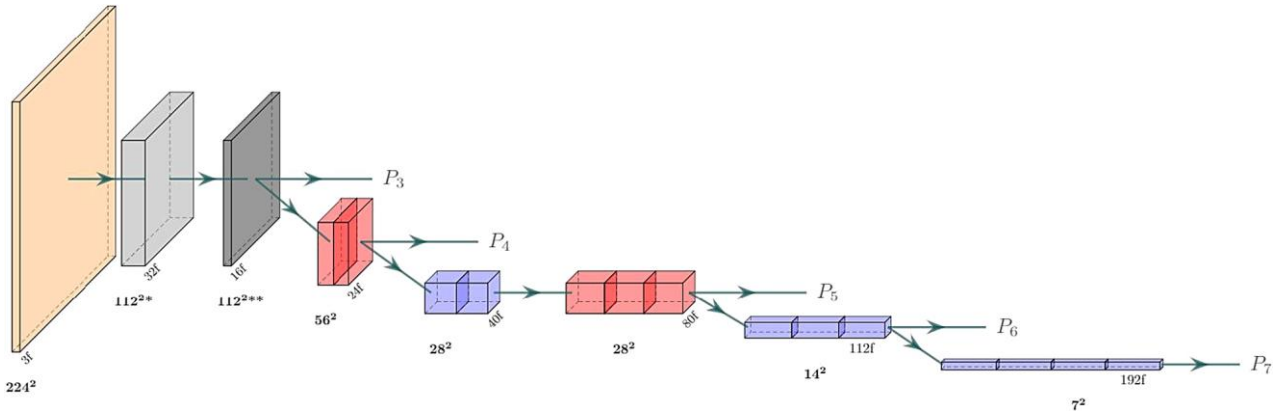


Figure 6. Architecture of EfficientNetB0

3-3-1-3- InceptionResNetV2

InceptionResNetV2, proposed by Szegedy et al. [13], represents the state-of-the-art deep learning architecture by realizing the powers of both an Inception module and Residual Network (ResNet). With the incorporation of multiscale feature extraction and residual learning, InceptionResNetV2 does both-fast and accurate. Detection of keratoconus can be such a medical image processing task where identification of local and global patterns may play a vital role in accurate classification.

The architecture of InceptionResNetV2 provides an Inception module as the core feature using parallel convolutional layers with kernel sizes of 1×11, 3×33, and 5×55 to attain higher-dimensional features with finer details and larger context information simultaneously. Thus, this fits perfectly for analyzing such complex corneal structures. And among the other innovative features in InceptionResNetV2 include residual connections inspired by ResNet, enabling the network to learn residual mappings instead of direct mappings, which makes the optimization process easier and reduces the vanishing gradient problem. This allows the training of deeper networks with efficiency, hence allowing the model to learn intricate patterns in data. However, InceptionResNetV2 has reduction blocks that downsample the feature maps while preserving all the important information to manage its deep architecture computationally. In this way, the network can remain computationally efficient when processing high-resolution medical images. Batch normalization after every convolution helps to stabilize learning and accelerate convergence in training.

Architecture of InceptionResNetV2

InceptionResNetV2 shown in Figure 7 [54] employs a modular structure that includes several critical components:

- Stem Block: This initial block preprocesses input images, reducing spatial dimensions while extracting foundational features. It prepares the images for deeper processing in the subsequent blocks.
- Inception-ResNet Blocks: These are the core feature extraction units of the architecture. They combine the multi-scale capabilities of Inception modules with the optimization benefits of residual connections. There are three types of Inception-ResNet blocks (A, B, and C), each designed to extract features at different levels of abstraction.
- Reduction Blocks: These blocks downsample the feature maps and reduce computational overhead while preserving important patterns. Reduction-A and Reduction-B are strategically placed between Inception-ResNet blocks.
- Global Average Pooling: This layer aggregates spatial information from the feature maps, reducing their dimensionality while retaining critical details for classification.
- Fully Connected Layer: The final layer converts the extracted features into a compact representation, making them ready for classification.

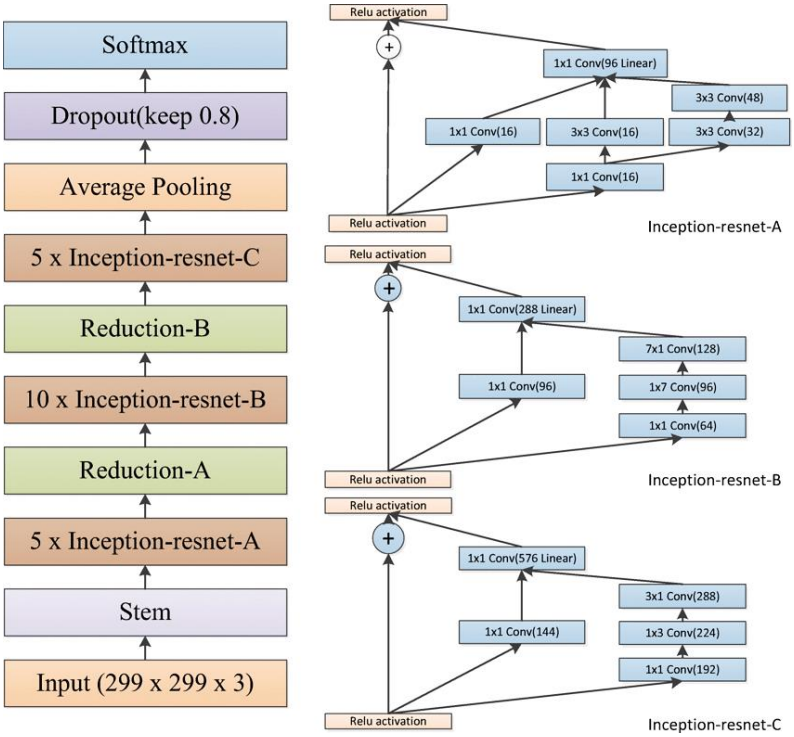


Figure 7. Architecture of InceptionResNetV2

3-3-1-4- InceptionV3

InceptionV3, proposed by Szegedy et al. [14], is a very powerful and efficient CNN for feature extraction, an improved version of its earlier versions. It adopts several advanced techniques to attain high accuracy with reduced computational complexity, including factorized convolutions, label smoothing, and auxiliary classifiers. These innovations make the InceptionV3 suitable for medical image classification problems like keratoconus detection. The architecture of InceptionV3 is shown in Figure 8 [55].

The most salient feature of the InceptionV3 is its reliance on factorized convolutions. Rather than using a big convolutional filter, say 5×55, it uses a smaller and sequential one: two 3×33 convolution operations. This saves on computational cost but still preserves the model's capability to learn complex patterns. Another approach it takes for efficiency optimization is asymmetric convolutions: a 3×33 operation is decomposed into 1×31 followed by a 3×13 convolution.

InceptionV3 uses label smoothing during training to soften the true labels in order to improve generalization and reduce overfitting. This prevents the model from being too confident in its predictions and hence makes it more robust for noisy data. Another novelty is the usage of auxiliary classifiers, i.e., some middle layers where extra supervision will be performed. These classifiers act like regularizers but, more importantly, they enhance the gradient flow, which for deeper networks would impede the optimization because of the diminishing gradients.

It also embeds some very effective grid reduction techniques that downsample the feature maps without losing the essential details. This ensures that the InceptionV3 network captures critical patterns efficiently for a task such as detection of keratoconus.

Inception V3 is a modular model consisting of many performance-optimized components. Its stem block, present at the beginning, does resizing of the input image and extracts low-level features. In the core of the network, several Inception modules process features in parallel at multiple scales using convolution with kernel sizes 1×11, 3×33, and 5×55, thus making it possible to catch both the fine details and more general patterns of the network.

The reduction blocks between the Inception modules serve the purpose of keeping computational costs in check: they downsample the feature maps, reducing the spatial dimensions and retaining important information. A global average pooling layer at the end of the network aggregates the spatial information, with a reduction in the number of parameters and overfitting risks. The last fully connected layer outputs the features extracted, to be used either for classification or passed onto a hybrid model for further analysis.

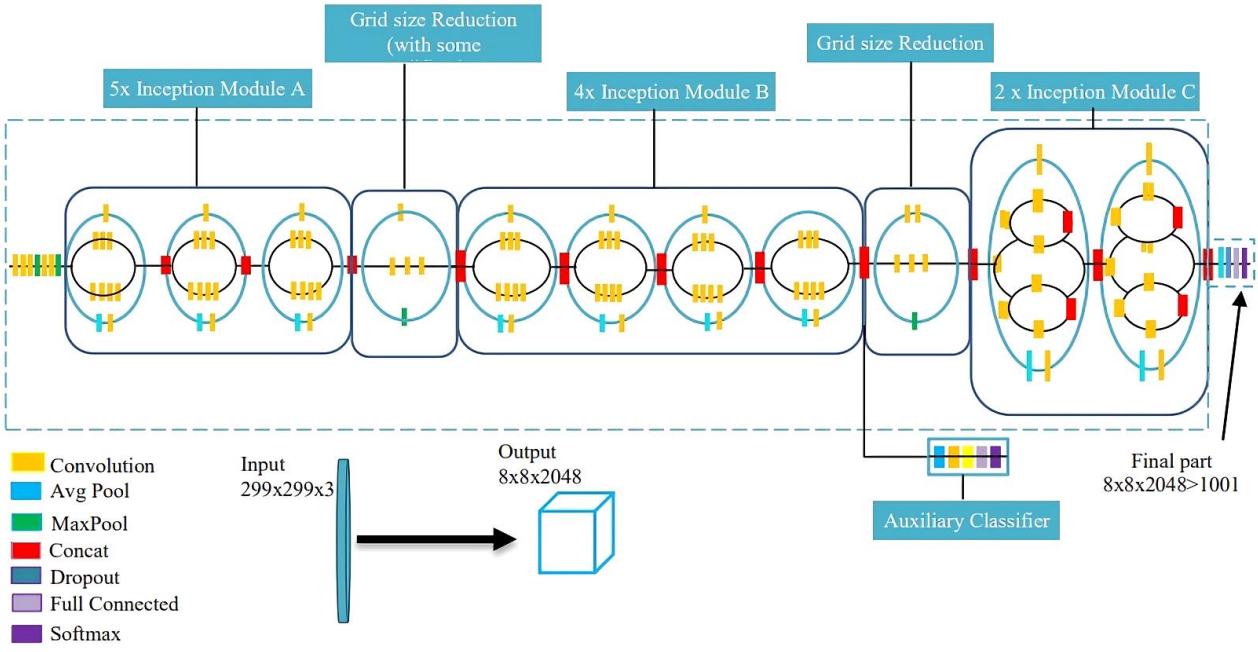


Figure 8. Architecture of InceptionV3

3-3-1-5- MobileNetV2

MobileNetV2 is a very efficient and light convolutional neural network proposed by Sandler et al. [15], balancing accuracy and computational cost. Its main goal is to allow the execution of CNNs on resource-constrained mobile and embedded vision systems. It leverages inverted residual structures combined with linear bottlenecks, significantly reducing parameters and MACs while achieving state-of-the-art performance. That would make it an ideal feature extractor for medical imaging tasks such as the detection of keratoconus, where efficiency in resources is paramount. The architecture of MobileNetV2 is shown in Figure 9 [56].

MobileNetV2 mainly features depth-wise separable convolutions that factorize a standard convolution into two separate operations: depth-wise convolution and pointwise convolution. The dept-wise convolution applies a single filter to each input channel, performing spatial filtering, while the pointwise convolution with 1×11 filters combine the outputs of the depth-wise convolution across channels. This factorization reduces the computation compared to standard convolution by a big margin.

In this context, MobileNetV2 also proposes a new structure, called inverted residuals, for which residual connections are applied to bottleneck layers. Instead of connecting a wide, high-dimensional layer, MobileNetV2 connects thinner, lower-dimensional bottleneck layers. Each bottleneck block consists of three stages: expansion, depth-wise convolution, and projection. Expansion increases the dimensions of the input, depth-wise convolution extracts the spatial features, and in the projection phase, the dimensions are shrunk to the size of the bottleneck. This structure provides much more efficient computation while retaining the important features.

Another novelty in MobileNetV2 is the inclusion of linear bottlenecks in place of non-linear activations. Though most of the CNN models use ReLU activations, these might destroy the information in a low-dimensional space; therefore, the bottlenecks of MobileNetV2 consist of a linear activation function. This will help in not losing some crucial information while enhancing feature extraction, particularly in complex datasets like medical images.

The MobileNetV2 architecture starts with a standard convolution layer followed by a sequence of bottleneck residual blocks, ends with a global average pooling layer and a fully connected layer. This would include a kernel size of 3 × 33, with 32 filters in the initial convolution layer that moves with a stride of 2 for the reduction of spatial dimensions of the input for the extraction of rudimentary features, followed by blocks of bottleneck residual units implementing the major feature extraction. The block, in general, would also contain an expansion phase, depth-wise convolution, and projection phase. In blocks, residual connections exist wherever the dimensions of input and output agree.

The network is ended with a global average pooling layer that aggregates the spatial information of the feature maps, reducing each map to a single value. This reduces the number of parameters and consequently minimizes overfitting. This finally generates a compact feature representation through a fully connected layer that can then be used for classification tasks or as input to a hybrid model like the Vision Transformer.

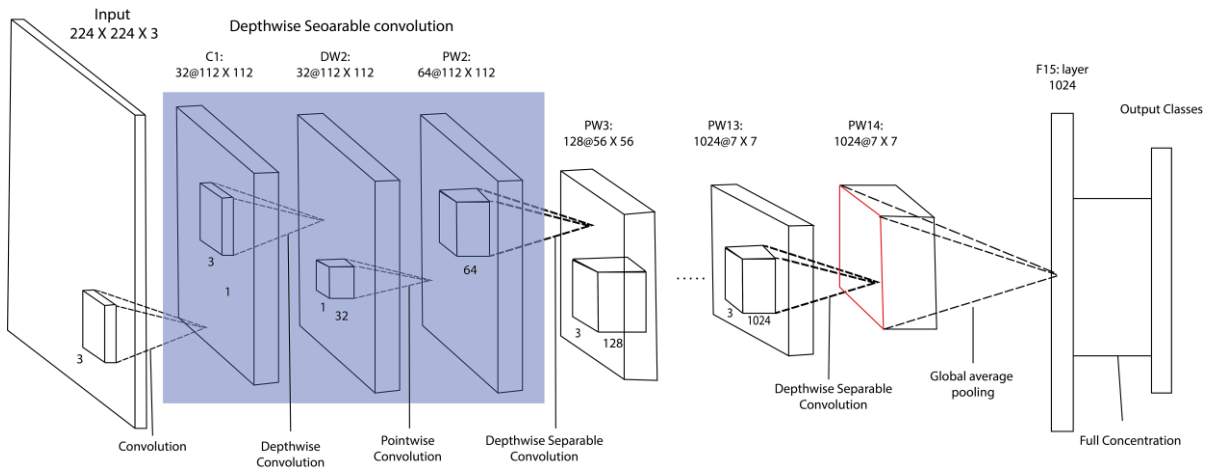


Figure 9. Architecture of MobileNetV2

3-3-1-6- ResNet50

ResNet50 is a deep convolutional neural network proposed by He et al. [12], and it overcomes difficulties inherent in the training of very deep networks due to the vanishing gradient problem; its novelty was residual learning, thus enabling efficiency in training by skipping information across one or more layers through its skip connections. With 50 layers, ResNet50 is adopted for many fields in feature extraction, including that of medical imaging. It has proved efficient for learning of hierarchical representation; therefore, this model could be very efficient for such a complicated task as keratoconus detection. The architecture of ResNet50 is shown in Figure 10 [57].

It brings the main novelty to ResNet50 with residual blocks, which are helpful in allowing the network to learn residual mappings rather than direct mappings. Every residual block will have a skip connection to add the input of the block directly to its output. Thus, this design avoids vanishing gradient problems and further improves the flow of gradients at the backpropagation; hence it allows the training of very deep networks.

Another important feature in the architecture is the bottleneck design inside the residual blocks. Each block contains three layers: one 1×11 convolution for reducing the channel number, one 3×33 convolution for spatial filtering, and another 1×11 convolution for restoring the dimensions. This bottleneck design reduces computational complexity while preserving the capability of the model for high-level feature capture.

ResNet50 also implements batch normalization to normalize the feature maps after every convolution. This helps in stabilizing and speeding up the training by reducing internal covariate shifts. Also, ResNet50 uses global average pooling in the last few layers, which aggregates spatial information in a way that reduces the total number of parameters, hence reducing overfitting and enhancing generalization.

ResNet50 has a design of the architecture that involves one convolutional layer first; after that comes residual blocks divided into four stages and at the end of global average pooling and finally the fully connected output layer.

The first layer is the convolution of 7×77 with 64 filters, a stride of 2; it is then followed by a max pooling layer-a combination that reduces spatial dimensions while extracting low-level features. This is followed by passing four stages of residual blocks, each consisting of several bottleneck layers:

- Stage 1: Contains 3 residual blocks.
- Stage 2: Contains 4 residual blocks.
- Stage 3: Contains 6 residual blocks.
- Stage 4: Contains 3 residual blocks.

In each residual block, skip connections concatenate the input to the output of these convolutional layers, which ensure that gradients backpropagate very efficiently and that local features are persisted. The network concludes with a Global Average Pooling layer, to summarize spatial information from feature maps into a small representation, and the fully connected one for classification purposes.

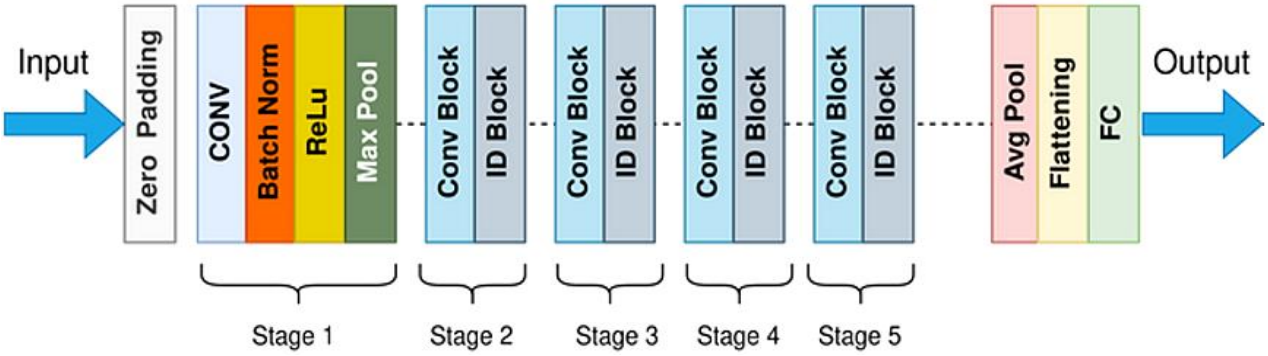


Figure 10. Architecture ResNet50

3-3-1-7- VGG16

VGG16 is a seminal CNN, proposed by Simonyan & Zisserman [16]; it is essentially marked by simplicity and depth. The network comprises 16 layers, out of which 13 are convolutional layers and the rest are fully connected, each having an architectural structure in common. In VGG16, the small 3×33 filters make it very efficient in feature extraction throughout the network. It can capture low-level information, complex patterns, and is very useful in hierarchical structure for the detection of keratoconus in medical imaging. The architecture of VGG16 is shown in Figure 11 [58].

The core of VGG16 is the same to use 3×33 convolutional filters, which are the smallest size of filters that can show the spatial relationships in an image. This is done in such a way that the process of feature extraction will be consistent in all layers. This leads to a high degree of accuracy with simplified architecture. This depth allows for hierarchical feature extraction, where early layers capture simple patterns such as edges and textures, and later layers combine those into more complex structures and global patterns.

It basically implements max pooling after every few convolutional layers in VGG16 to handle the growth in parameters that would be present in deeper layers. This reduces the spatial dimensions of the feature maps but keeps most of the vital information intact in a very computationally friendly way. In the end, this network used a fully connected layer in which dense connectivity links all the features extracted before final predictions are made. The use of ReLU activation at each layer introduces non-linearity, hence enabling the network to learn even the most complicated mapping between input images and output classes.

The architecture of VGG16 is designed as a series of convolutional layers grouped into five blocks, each followed by a max pooling layer, and concludes with three fully connected layers.

- Input Layer: The input images are resized to 224×224 pixels, which is the standard input size for VGG16. This ensures uniformity and compatibility across all images.
- Convolutional Layers: The network consists of 13 convolutional layers grouped into five blocks:
  - Block 1: Two 3×33 convolutional layers with 64 filters each, followed by max pooling.
  - Block 2: Two 3×33 convolutional layers with 128 filters each, followed by max pooling.
  - Block 3: Three 3×33 convolutional layers with 256 filters each, followed by max pooling.
  - Block 4: Three 3×33 convolutional layers with 512 filters each, followed by max pooling.
  - Block 5: Three 3×33 convolutional layers with 512 filters each, followed by max pooling.
- Fully Connected Layers: After flattening the output of the final max pooling layer, the network passes the data through three fully connected layers. The last fully connected layer uses a SoftMax activation function to produce class probabilities.
- Max Pooling: Max pooling, applied with a 22 × 2 kernel and a stride of 2, reduces the dimensions of the feature maps at regular intervals, improving computational efficiency without losing critical features.

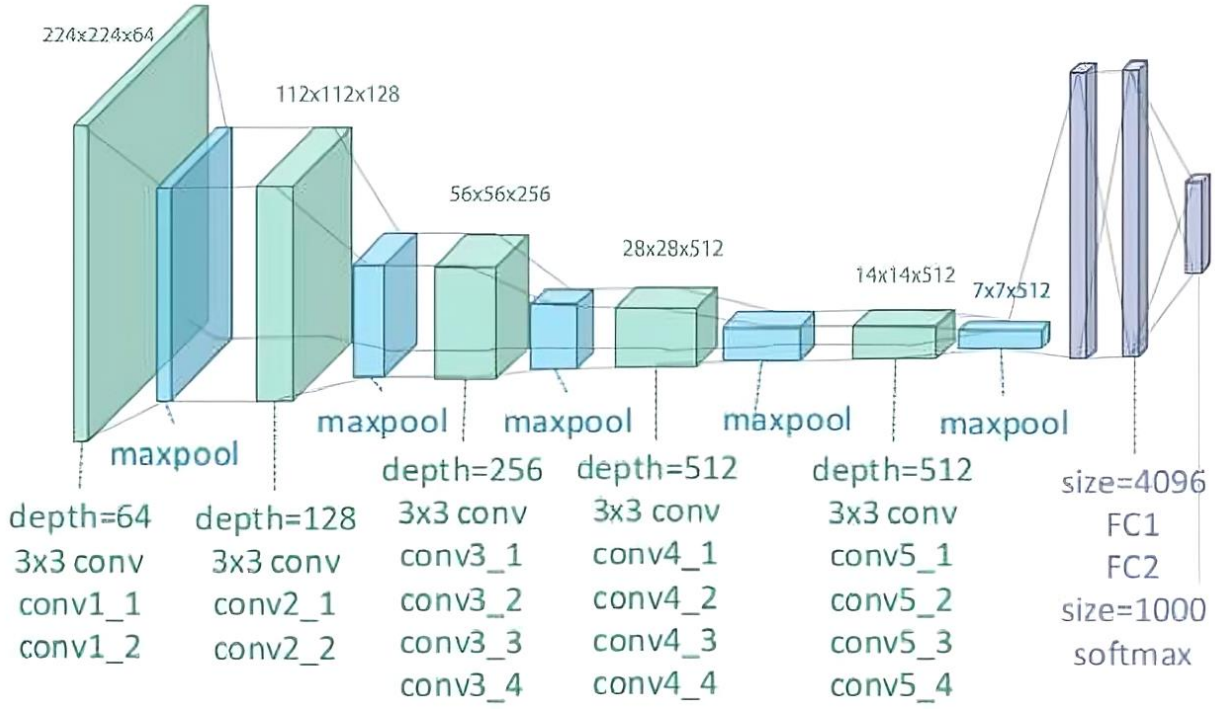


Figure 11. Architecture of VGG16

3-3-1-8- VGG19

VGG19 is an extended form of VGG16 architecture and proposed by Simonyan & Zisserman in 2015 [16]. It consists of 19 layers, which are 16 convolutional layers and 3 fully connected layers. The increased depth in VGG19 allows it to capture complex and fine details, very useful in tasks where minute details need to be analyzed. Its hierarchical structure allows the modeling of patterns from simple edges to complex structures, which is important in applications such as keratoconus detection in medical imaging. The architecture of VGG19 is shown in Figure 12 [59].

Thus, the defining features of VGG19 include that all convolutional layers are defined using 33×3 convolution filters consistently. These filters use a stride of 1, combined with padding, to work out the spatial relationships in input images. By having more convolutional layers than the VGG16 model, it is capable of learning a greater level of detail and more fine-grained information, especially as it goes further into the model.

The max pooling layers are placed after small groups of convolutional layers to downsample the feature maps in space, allowing computational efficiency while retaining all the important information. The network is terminated by three fully connected layers that summarize and interpret the features extracted to get the final classification. Throughout the convolutional and fully connected layers, the ReLU activation function is used, introducing non-linearity to let the network learn complex mappings from input to output.

The VGG19 architecture contains five convolutional blocks, each one followed by max pooling, followed by fully connected layers at the end.

- Input Layer: The network accepts input images resized to 224×224 pixels, standardizing the input dimensions for uniform processing.
- Convolutional Layers: The convolutional layers are organized into five blocks:
  - Block 1: Two 3×33 convolutional layers with 64 filters each, followed by max pooling.
  - Block 2: Two 3×33 convolutional layers with 128 filters each, followed by max pooling.
  - Block 3: Four 3×33 convolutional layers with 256 filters each, followed by max pooling.
  - Block 4: Four 3×33 convolutional layers with 512 filters each, followed by max pooling.
  - Block 5: Four 3×33 convolutional layers with 512 filters each, followed by max pooling.
- Fully Connected Layers: After the final max pooling layer, the output is flattened and passed through three fully connected layers. The last layer uses a SoftMax activation function to output class probabilities.
- Max Pooling: Each max pooling layer uses a 22×2 kernel with a stride of 2, reducing the spatial dimensions of the feature maps and focusing on essential features.

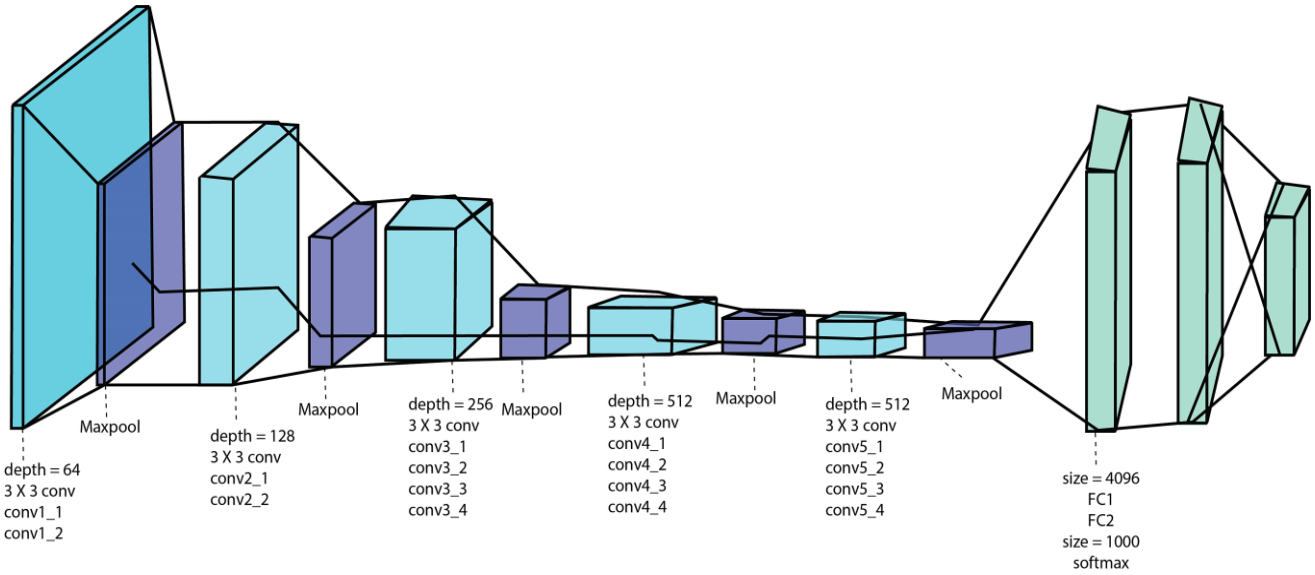


Figure 12. Architecture of VGG19

3-3-2- Model Building

It was finally implemented in this work, the model that merged features from pretrained deep learning models to a classification model based on the Vision Transformer. This method fully utilizes the local-to-global hierarchical feature learning in CNNs with the global attention mechanism of transformers to achieve accurate and robust keratoconus detection. The proposed model is designed in a way to transcend the limitations inherent in single standalone models through effectively embedding diverse feature representation with strong classification potential.

3-3-2-1- Feature extraction with Pretrained Models

For constructing the fusion model, feature extraction with any of the pre-trained transfer models is carried out as the initial step. Features from various deep architectures such as DenseNet121, EfficientNetB0, InceptionResNetV2, InceptionV3, MobileNetV2, ResNet50, VGG16, and VGG19 are considered in this work. These pre-trained models, while being trained with huge datasets such as ImageNet, can capture hierarchical features in images. A number of features are extracted from input data using:

Low-Level Features: In CNN shallow layers, features like edges, textures, and primitives are extracted.

Intermediate Features: Middle layers capture patterns that indicate structural irregularities, such as variation in curvature or thickness of the cornea.

High-Level Features: Complex global patterns representing the characteristics of keratoconus are extracted in deeper layers.

The output from these pre-trained models consists of a set of rich feature maps, representing the input images in high-dimensional space. These features are compact, descriptive, and ready for further analysis.

3-3-2-2- Feature Aggregation

After feature extraction, features are flattened or pooled to get the fixed-length vector representation, which is a must to make it compatible with the Vision Transformer for sequential input data. Global average pooling or flattening is generally used to reduce the feature maps without losing important information.

Global Average Pooling: The aggregation of the spatial information across each feature map.

Flattening: It changes the multi-dimensional feature maps to one-dimensional vector form.

This aggregated feature vector retains major patterns of the original image, yet with a reduction in its dimensionality, making it apt for input into the transformer.

3-3-2-3- Classification with Vision Transformer Embedded Fusion Model

The proposed fusion model for the classification of keratoconus disease integrates the local feature extraction capability of the pre-trained CNNs with the global contextual understanding provided by ViT. This hybrid architecture leverages complementary strengths of both models to achieve robust performance in medical image classification tasks.

The pre-processing of the dataset, therefore, guarantees that the input images are standardized to be compatible with the deep learning models. That is, the pre-processing steps include resizing the images into uniform dimensions, normalizing the pixel values within the range of [0,1] [0,1], and converting the images into tensors. All of these indeed prepare the data for analysis and enhance computational efficiency during training.

The feature extraction stage of the proposed fusion model relies on several pre-trained CNNs including DenseNet121, EfficientNetB0, InceptionResNetV2, InceptionV3, MobileNetV2, ResNet50, VGG16, and VGG19. These deep architectures will be initialized with pre-trained weights on the ImageNet dataset, enabling them to learn complex local patterns and their spatial hierarchies from the input images. Each of the CNNs processes the image through convolutional layers that extract localized features, including edges and textures, pooling layers that reduce the spatial dimensions, and fully connected layers that produce high-dimensional feature vectors. These feature vectors capture detailed local information, which is crucial for correct classification.

These feature vectors are fed to the Vision Transformer, which constitutes the backbone of the fusion model. Specifically, the feature vector of each CNN is segmented or divided into smaller patches, further flattened and projected to a fixed dimensionality for extracting features as a sequence of patches by the ViT in learning global dependencies and contextual relations along the whole image. These embeddings are further added with positional encodings to retain the spatial arrangement of the patches. ViT uses a multi-head self-attention mechanism that calculates the importance of each patch with respect to others, capturing both short-range and long-range dependencies. The output of the attention mechanism is further refined through feedforward networks, enabling the model to learn robust representations of the data. The crucial elements of ViT, as applied in the model proposed, are explained in detail below.

Input Embedding

The first step in the Vision Transformer involves preparing the feature vector obtained from CNN for processing by the transformer. This requires tokenization and positional encoding.

Feature Tokenization

The high-dimensional feature vector output from the CNN is divided into fixed-sized segments, referred to as tokens.

Each token is represented as a vector of fixed size *d*, forming a sequence of tokens $X = x_1, x_2, \ldots . x_N$, where *N* is the number of tokens.

Positional Encoding

Transformers lack inherent knowledge of the spatial arrangement of tokens, which is critical for image data. To address this, positional encoding is added to each token.

Positional encoding captures the spatial relationship between tokens and is defined as:

$$PE(p, 2i) = \sin\left(\frac{p}{1000^{\frac{2i}{d}}}\right), PE(p, 2i+1) = \cos\left(\frac{p}{1000^{\frac{2i}{d}}}\right) \quad (1)$$

here: *p*: Position of the token in the sequence; *i*: Index of the feature dimension; *d*: Dimensionality of the token embedding.

The positional encoding is added to the token embeddings:

$$Z_0 = X + PE \quad (2)$$

This ensures the transformer is aware of the spatial relationships among the tokens, enabling it to preserve the positional context of image features.

Multi-Head Self-Attention

The core of the Vision Transformer is its multi-head self-attention mechanism, which enables the model to capture relationships between tokens, both locally and globally.

Self-Attention Mechanism

Self-attention calculates the relevance of each token to every other token in the sequence. For a token x_i, its attention scores relative to another token x_j is computed as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

Here, $Q = ZW_Q$: Query matrix obtained by projecting the input Z using the weight matrix W_Q; $K = ZW_K$: Key matrix obtained by projecting Z using W_K; $V = ZW_V$: Value matrix obtained by projecting Z using W_V; d_k: Dimensionality of the keys (a scaling factor to stabilize gradients).

The attention mechanism computes a weighted sum of the values *V*, where the weights are derived from the similarity between queries *Q* and keys *K*.

Multi-Head Attention

Instead of a single attention mechanism, ViT uses multi-head attention to capture diverse relationships between tokens. Each head independently computes attention using different projection matrices (W_Q^h, W_K^h, W_V^h):

$$MultiHead(Q, K, V) = Concat(head_1, \dots\dots\dots, head_h)W_o \quad (4)$$

Here; $head_h = Attention(QW_Q^h, KW_K^h, VW_V^h$; W_o: Weight matrix for combining the outputs of all heads.

Multi-head attention enables the model to focus on different aspects of the input, such as local details and global patterns.

Feedforward Layers

After the self-attention mechanism, the output is passed through feedforward layers to refine the representations. The feedforward layer is applied to each token independently and consists of two linear transformations with a non-linear activation function (GeLU) in between:

$$FNN(x) = W_2 GeLU(W_1 x + b_1) + b_2 \quad (5)$$

here, W_1, W_2: Weight matrices of the feedforward layers; b_1, b_2: Bias terms; *GeLU(x)*: Gaussian Error Linear Unit activation function.

The feedforward layers improve the representational capacity of the model, enabling it to refine the learned dependencies and capture higher-level abstractions.

Output Layer

The final output of the Vision Transformer is processed by a classification head to map the learned representations to the target classes.

Classification Head

The class token, typically prepended to the input sequence during the embedding step, aggregates the global context learned by the transformer. This token is passed through a fully connected layer to generate the logits:

$$y = W_c \cdot z_{class} + b_c \quad (6)$$

here, W_c: Weight matrix of the classification layer; b_c: Bias term; z_{class}: Final embedding of the class token.

SoftMax Activation

The logits *y* is passed through a SoftMax activation function to convert them into probabilities:

$$p_i = \frac{e^{y_i}}{\sum_j e^{y_j}} \quad (7)$$

here: p_i: Probability of class *I*; y_i: Logit for class *I*; The class with the highest probability is selected as the predicted output.

3-3-3- Training Configuration

The training configuration of the hybrid model includes critical elements that guide how the model learns from the data. These elements include the loss function, optimizer, and key training parameters such as batch size and the number of epochs. Below is a detailed explanation of each component, including their equations and significance.

3-3-3-1- Loss Function: Cross Entropy Loss

Cross Entropy Loss is used to calculate prediction errors in multi-class classification tasks. It measures the difference between the predicted probability distribution and the true class labels. For a dataset with *N* samples, *C* classes, and the true class labels represented as one-hot encoded vectors, the Cross Entropy Loss is defined as:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} log(\hat{y}_{i,c}) \quad (8)$$

here: $y_{i,c}$: Ground truth (true label) for the *i*-th sample and *c*-th class. It is 1 if the *c*-th class is the true label for the *i*-th sample; otherwise, it is 0; $\hat{y}_{i,c}$: Predicted probability for the *c*-th class of the *i*-th sample (output of the SoftMax layer); *N*: Total number of samples in the batch.

3-3-3-2- Optimizer: AdamW

AdamW (Adam with Weight Decay) is used to update model weights during training. It extends the original Adam optimizer by incorporating weight decay to prevent overfitting and improve generalization. Adam combines the advantages of two popular optimization methods:

Momentum: Incorporates the exponentially decaying average of past gradients.

Adaptive Learning Rate: Scales the learning rate for each parameter based on the historical gradient magnitude.

The parameter update rule for Adam is:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \tag{9}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \tag{10}$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \hat{v}_t = \frac{v_t}{1-\beta_2^t} \tag{11}$$

$$\theta_t = \theta_{t-1} - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon} \tag{12}$$

here: g_t: Gradient of the loss with respect to parameter θ_t at step *t*; m_t, v_t: First and second moments of the gradients; β_1, β_2: Exponential decay rates for m_t and v_t, typically set to 0.9 and 0.999, respectively; η: Learning rate; ϵ: Small constant to prevent division by zero.

AdamW incorporates weight decay (λ) into the parameter updates, which penalizes large weights to prevent overfitting:

$$\theta_t = \theta_{t-1} - \eta \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon} + \lambda \theta_{t-1} \right) \tag{13}$$

Here; $\lambda\theta_{t-1}$: Regularization term that encourages smaller weights.

3-3-3-3- Training Parameters

Batch Size: A batch size of 32 is selected here to reach an efficient balance between computational efficiency and learning stability. Smaller batches allow for quicker updates of weights and help escape shallow local minima in the loss landscape due to noisier estimates of the gradient. However, they are more computationally wasteful because they provide less accurate gradient estimation, which requires smaller learning rates to maintain stability. This batch size represents an important practical compromise that guarantees dynamic learning without overloading computation resources.

Epochs: The model trains for 10 epochs, each being a full pass over the dataset. This allows for enough iteration for the model to learn patterns and converge without leading to overfitting. Frequent monitoring of validation loss during training ensures early stopping if performance stabilizes or starts to degrade, hence keeping generalization optimal. Together, these parameters enable strong training for the keratoconus detection task.

3-4- Model Evaluation

To investigate thoroughly the performance of this hybrid model in the context of keratoconus detection, several metrics are being used: accuracy, precision, recall, and F1-score. These categorical performance metrics can help to understand how capable the model will be for multiclass classification problems and also when the data is highly imbalanced. Each metric will be described thoroughly herein.

3-4-1- Accuracy

Accuracy is the most straightforward evaluation metric, measuring the proportion of correct predictions out of the total predictions made by the model. It provides a general measure of the model's overall performance and is calculated by the following formula:

$$Accuracy = \frac{number\ of\ correct\ prediction}{Total\ number\ of\ prediction} = \frac{TP+TN}{TP+TN+FP+FN} \tag{14}$$

where: TP (True Positive): Correctly predicted positive samples; TN (True Negative): Correctly predicted negative samples; FP (False Positive): Incorrectly predicted positive samples; FN (False Negative): Incorrectly predicted negative samples.

3-4-2- Precision

Precision focuses on the quality of the positive predictions made by the model. It evaluates the proportion of true positive predictions among all samples predicted as positive and is calculated with the following formula:

$$Precision = \frac{TP}{TP+FP} \quad (15)$$

3-4-3- Recall

Recall, also known as sensitivity or true positive rate, measures the model's ability to correctly identify all relevant instances in a given class. It reflects the proportion of true positive predictions relative to the total actual positive samples. Recall is calculated by the following formula:

$$Recall = \frac{TP}{TP+FN} \quad (16)$$

3-4-3- F1-Score

The F1-score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance. It is particularly useful when dealing with imbalanced datasets, as it combines both precision and recall into a single metric. F1-score is calculated by the following formula:

$$F1 - Score = 2 * \frac{precision*Recall}{Precision+Recall} \quad (17)$$

4- Result Analysis

All the pre-trained models were first trained and tested on the dataset, after which the hybrid models were trained and tested by combining feature extraction using those pre-trained models with the Vision Transformer for classification. The Vision Transformer alone was also conducted on the dataset to directly compare it with the standalone ViT and hybrid models. These are summarized in the following table, which gives an overview of the training and validation accuracies for both the pre-trained and hybrid models. This provides a broader comparison of the overall performance of these hybrid models outperforming their corresponding pre-trained model baselines and competitive performance among the standalone baselines of a Vision Transformer model.

4-1- Discussion of Performances of the Hybrid Models

The results of all the pretrained and hybrid models combined with the ViT are highlighted for training, validation, and testing accuracies in Table 1. The hybrid models seem to perform a little better compared to the pretrained models across all these metrics. The standalone application of the Vision Transformer also turns out quite well, while hybrid models provide small yet significant improvements by combining CNN-based feature extraction with the global attention mechanism of ViT.

The single DenseNet121 pretrained model has a performance of 64.83% in training and 69% testing accuracy, with its hybrid reaching a significant increase to 99.28%, which is increased by 30.28%; therefore, it could be proved that the dense connectivity of DenseNet121 benefits significantly from ViT about the global contextual information to get a stronger classification performance. Meanwhile, the same EfficientNetB0 and InceptionResNetV2, performing modestly with 66% and 67%, respectively, will see notable improvements in their hybrid forms. They achieved a test accuracy of 98.84% under the hybrid approach when ViT would make up for deficiencies in the capability of non-trivial dependencies' representations.

InceptionV3 performs well as a standalone model, achieving a testing accuracy of 90%. Its hybrid version shows a slight improvement, reaching a testing accuracy of 99.06%. This modest gain reflects InceptionV3's strong standalone feature extraction capabilities, with the hybrid model further refining global relationships between features. MobileNetV2, another highly performing pretrained model with a testing accuracy of 98%, sees only marginal improvement in its hybrid form, achieving 98.9%. This suggests that MobileNetV2's lightweight architecture already extracts highly effective features, leaving less room for improvement when integrated with ViT.

ResNet50, on the other hand, shows the most significant improvement among all models when hybridized with ViT. Its standalone performance is relatively poor, with a testing accuracy of 66%, but the hybrid version achieves a remarkable testing accuracy of 99.18%, representing an improvement of 33.18%. This highlights how ViT's attention mechanism enhances ResNet50's limited representational power, enabling the hybrid model to better understand global dependencies in the dataset. VGG16 and VGG19 also benefit from hybridization, with VGG16 achieving a testing

accuracy of 99.28% and VGG19 reaching 98.47%. The improvement is more pronounced for VGG16, as its simpler architecture integrates well with ViT. In contrast, VGG19's deeper architecture appears to introduce redundancies or overfitting, which slightly limits its hybrid performance compared to the standalone ViT.

Table 1. Comparative performances of the models

	Pretrained Models			Hybrid with ViT		
Models	**Training Accuracy**	**Validation Accuracy**	**Testing Accuracy**	**Training Accuracy**	**Validation Accuracy**	**Testing Accuracy**
DenseNet121	64.83	72	69	99.11	99.28	99.28
EfficientNetB0	57.2	66.98	66	98.97	98.85	98.84
InceptionResNetV2	61.85	67.1	67	98.92	98.85	98.84
InceptionV3	87.35	91.95	90	99.1	99.06	99.06
MobileNetV2	97.89	98.13	98	99.19	98.91	98.9
ResNet50	56.21	66.48	66	99.17	99.19	99.18
VGG16	87.22	89.14	90	99.13	99.28	99.28
VGG19	61.94	65.61	68	99.06	98.47	98.47
ViT	99.05	98.63	98.62	-	-	-

The Vision Transformer alone achieves excellent performance, with a testing accuracy of 98.62%. However, the hybrid models generally outperform the standalone ViT by small margins. DenseNet121 and VGG16 hybrids show the largest improvements over ViT, reaching testing accuracies of 99.28%. These results underscore the benefits of combining the hierarchical feature extraction of CNNs with the global contextual understanding of ViT, even for a model as robust as the Vision Transformer (see Figure 13).

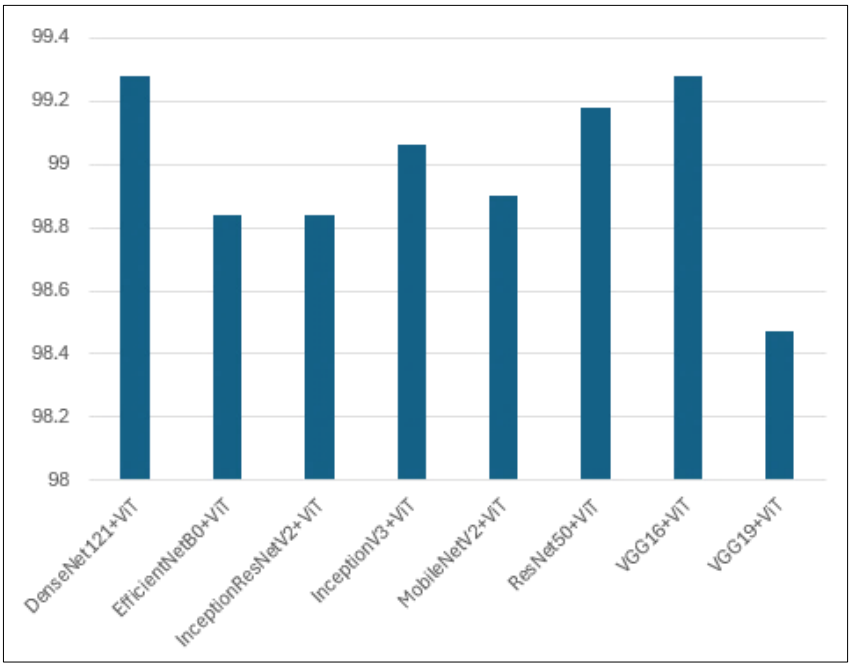


Figure 13. Comparison of performances of the hybrid models

Precision, recall, and F1-score were done to further analyze the performances of the models. The result shows that all hybrid models involving Vision Transformer (ViT) coupled with pre-trained CNNs put up very brilliant performances in classifying keratoconus into three classes, namely: Keratoconus, Normal, and Suspect. In fact, for all models, the classification of the Keratoconus class is outstanding, while several models obtained perfect Precision, Recall, and F1-Scores. Such is the case with DenseNet121 + ViT, ResNet50 + ViT, InceptionV3 + ViT, and VGG16 + ViT for the Keratoconus class, which all show perfect performances with an F1-score of 1.0 each, considering no false negatives or false positives regarding this class, and such regularity draws in the conclusion about how well the huge local feature extraction capability of CNN is mixed with the global attention mechanism of ViT. Table 2 gives a comparison of all the models using precision, recall, and F1-score.

Table 2. Comparison of precision, recall and F1-Score of each class

Models	Class	Precision	Recall	F1-Score
DenseNet121+ViT	Keratoconus	0.99	1	1
	Normal	0.99	0.99	0.99
	Suspect	0.99	0.99	0.99
EfficientNetB0+ViT	Keratoconus	0.99	1	0.99
	Normal	0.99	0.98	0.99
	Suspect	0.98	0.98	0.98
InceptionResNetV2+ViT	Keratoconus	1	1	1
	Normal	0.99	0.98	0.99
	Suspect	0.98	0.98	0.98
InceptionV3+ViT	Keratoconus	1	1	1
	Normal	0.98	0.99	0.99
	Suspect	0.99	0.98	0.98
MobileNetV2+ViT	Keratoconus	1	0.99	1
	Normal	0.98	0.99	0.99
	Suspect	0.99	0.98	0.98
ResNet50+ViT	Keratoconus	1	1	1
	Normal	0.99	1	0.99
	Suspect	0.99	0.98	0.99
VGG16+ViT	Keratoconus	1	1	1
	Normal	0.99	0.99	0.99
	Suspect	0.98	0.99	0.99
VGG19+ViT	Keratoconus	0.99	1	0.99
	Normal	0.98	0.98	0.98
	Suspect	0.98	0.97	0.97

Comparatively, there are slight changes between the performances of the models in the Normal and Suspect classes. For instance, EfficientNetB0+ViT and InceptionResNetV2+ViT show slight drops in Recall for the Normal class, with 0.98, showing misclassifications of true positives. For the Suspect class, a slight drop in Recall and F1-Score is similarly noted across these models. Although the difference is minimal, it suggests that distinguishing Suspect cases from the other two classes remains challenging due to their intermediate nature. Despite this, these models still maintain high Precision, which reflects their reliability in minimizing false positives.

This indeed presents quite strong results; among them are also InceptionV3 + ViT and MobileNetV2 + ViT. In particular, InceptionV3 + ViT has achieved perfect scores regarding the Keratoconus class and a close-to-perfect balance of Precision and Recall concerning the Normal and Suspect classes. MobileNetV2 + ViT, while maintaining high F1-Scores across all classes, has a slight drop in Recall for the Keratoconus class, 0.99, which means that a small fraction of the positive cases might not be detected. However, its overall performance remains robust, highlighting the model's efficiency despite its lightweight architecture.

Among these, ResNet50 + ViT has one of the best performances, achieving perfect Precision, Recall, and F1-Scores for the Keratoconus class while performing very well for the Normal class with a Recall of 1.0. The only slightly lower metric is in the class of Suspect, at Recall 0.98, as already showed from all the models in this trend. That proves how strong the balance in the performance for all classes is, as a consequence of combining deep features extracted by ResNet50 with global contextual learning made by ViT.

VGG16 + ViT also performs almost flawlessly, especially for the Keratoconus class, where it achieves a perfect F1-Score. The Normal and Suspect classes also show high Precision and Recall, indicating that the model can effectively distinguish between classes. On the other hand, VGG19 + ViT performs slightly less compared to other models, particularly in the Normal and Suspect classes. Precision and Recall for the Suspect class decrease to 0.98 and 0.97, respectively, providing an F1-Score of 0.97. The deeper architecture probably introduced redundancies or slight overfitting with VGG19, hence reducing its generalization capability across all classes.

4-2- Performance Visualization

In this section presents the training and validation accuracy of the hybrid models as well as the training and validation loss of the models. Finally, confusion matrices of the models will be presented to ensure the reliability of the models.

4-2-1-Accuracy and Loss Graph of DenseNet121+ViT

Figure 14 describes the accuracy and loss graph of DenseNet121+ViT. The plot presents the training and validation accuracy on the left and the loss on the right for the DenseNet121+ViT hybrid model for 10 epochs. The accuracy plot has a steep rise in the first few epochs, where both training and validation accuracy cross 95% at the 4th epoch and converge to around 99% at the 6th epoch. It is good learning and doesn't indicate any overfitting. Also, in the loss plot, there is a steep drop for both training and validation loss from roughly 0.6 and 0.4 to stabilize around 0.02 after the 4th epoch. This close alignment of the training and validation curve for accuracy and loss proves the stability of the model and its generalization capability. On the whole, the hybrid model.DenseNet121+ViT performs very well and is robust in terms of correct classification for the disease in question, keratoconus.

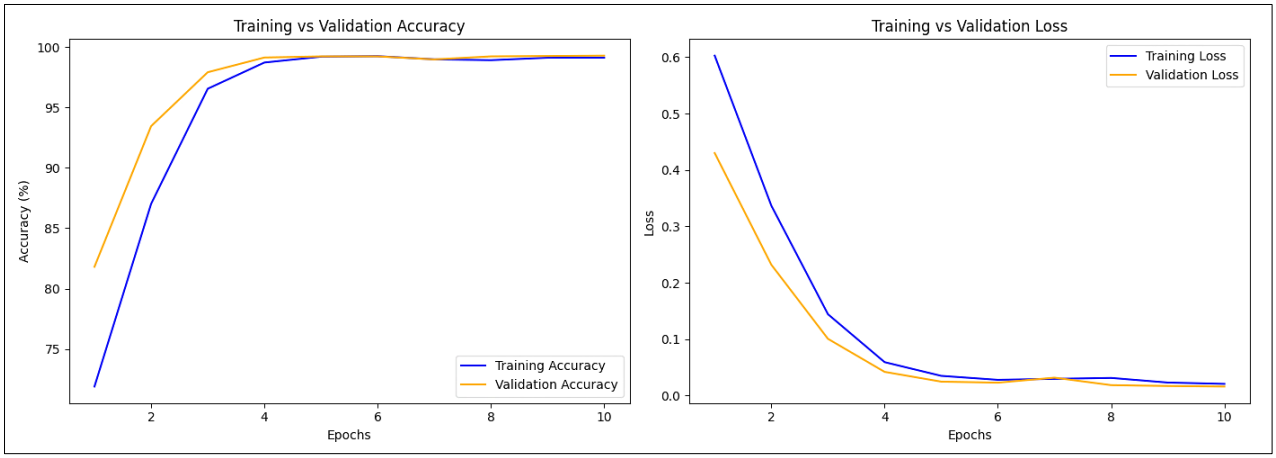


Figure 14. Accuracy and Loss graph of DenseNet121+ViT

4-2-2- Accuracy and Loss Graph of EfficientNetB0+ViT

Figure 15 illustrate the the training and validation accuracy (left) and loss (right) of the EfficientNetB0+ViT hybrid model over 10 epochs. The accuracy plot is quite smooth: the training and validation accuracy started at about 70% and 75%, respectively, and converged to approximately 99% at the 9th epoch. The two curves are also quite close throughout the process, which indicates that the model generalizes well. The loss plot presents a significant drop within the first few epochs from approximately 0.6 to near 0.01 around the 8th epoch, beyond which it stabilizes. Similarly, the validation loss follows a similar trend; however, a minor fluctuation around the 7th epoch is noticeable before convergence. In general, the proposed hybrid approach with EfficientNetB0+ViT has a very good learning capability that achieves almost complete accuracy with minimum loss, despite minor variations during training.

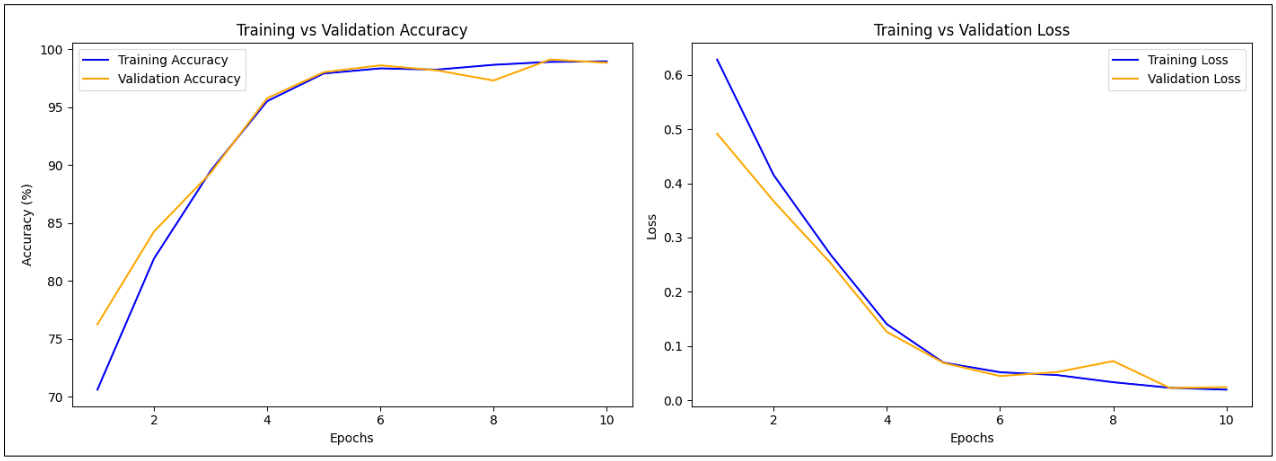


Figure 15. Accuracy and Loss graph of FfficientNetB0+ViT

4-2-3- Accuracy and Loss Graph of InceptionResNetV2+ViT

The Figure 16 shows training and validation accuracy (left) and loss (right) of the InceptionResNetV2+ViT hybrid model in 10 epochs. The accuracy plot shows a steep rise in training and validation accuracy, with the first epoch starting at roughly 74% training accuracy and 86% validation accuracy. At the 3rd epoch, both metrics go beyond 97% and then converge very close to 99% in the 5th epoch onward with a gentle trend of loss afterward. It shows a drastic decrease for training and validation loss at initial epochs. In addition, training loss decreased from around 0.58 to about 0.01 by the 4th epoch. The validation loss also has a similar trend, settling just above zero with minimal fluctuation, showing effective learning with excellent generalization. Overall, InceptionResNetV2+ViT has very outstanding performance: it converges fast, it is highly accurate, and it has small loss, making the model robust and reliable for the classification of keratoconus.

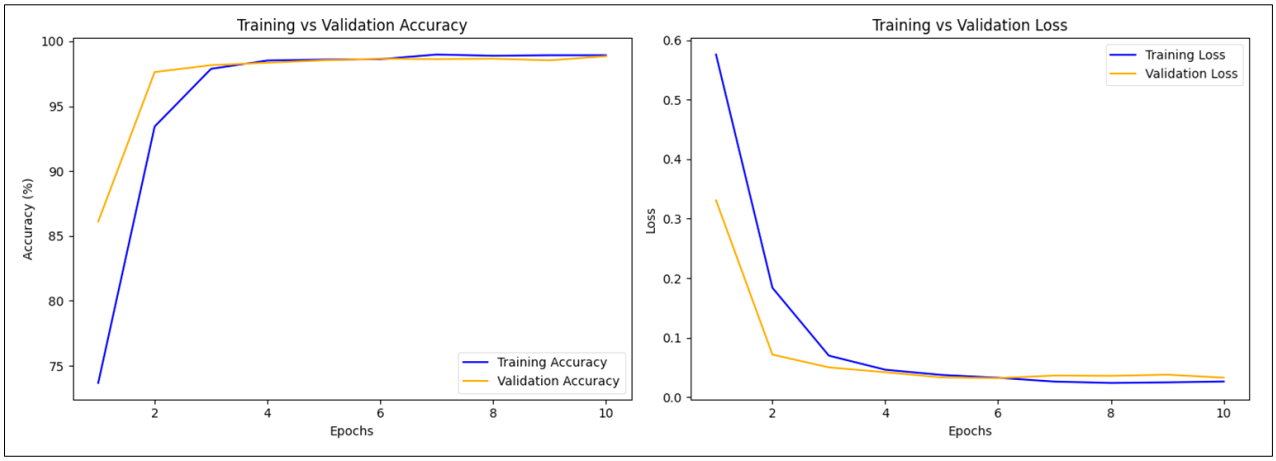


Figure 16. Accuracy and Loss graph of InceptionResNetV2+ViT

4-2-4- Accuracy and Loss Graph of InceptionV3+ViT

Figure 17 illustrates the training and validation accuracy (left) and loss (right) for the InceptionV3+ViT hybrid model during 10 epochs. The training and validation accuracy starts from around 70% and 80%, respectively, and reaches close to 99% within the 5th epoch with steep rises in the initial few epochs, while it remains stable during the rest of the training. They do follow a quite similar pattern in the graph between training and validation loss over the first few epochs, starting at 0.8 for training loss and 0.6 for validation loss and then both stabilizing at around 0.01 at the 6th epoch. Moreover, the high proximity between train and validation metrics shows good generalization and no noticeable overfitting. In a nutshell, inceptionResNetV2+ViT has very robust learning with high accuracy and low loss, hence highly reliable in the classification of keratoconus.

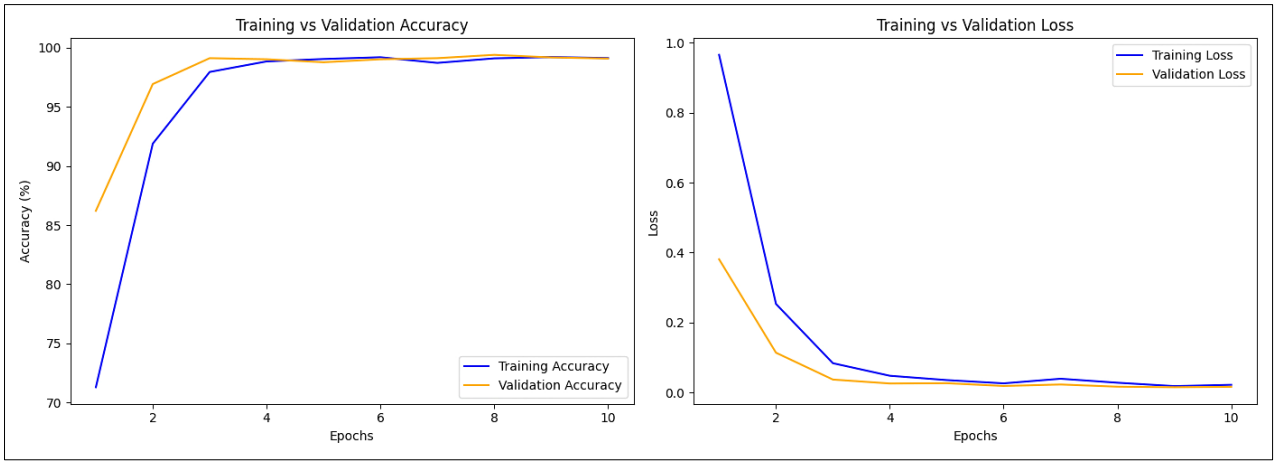


Figure 17. Accuracy and Loss graph of InceptionV3+ViT

4-2-5- Accuracy and Loss Graph of MobileNetV2+ViT

Figure 18 includes the training and validation accuracy and loss for a MobileNetV2+ViT hybrid model, trained for 10 epochs. Accuracy in this plot is growing steadily from approximately 70% to 80%, starting on and converging around 99% before the 5th epoch, with no further major variations. The same trend can be observed in the following graph:.

The loss plot indicates a big drop in the training and validation loss in the initial epochs where the training loss decreases from approximately 0.6, that of validation from 0.5, further stabilizing to nearly 0.01 in the 6th epoch. That it can create such similar curves of training and validation during epochs demonstrates very good generalization and stability by the model. Overall, the MobileNetV2+ViT hybrid model performs well, yielding high accuracy and low loss, hence being efficient and reliable for the classification of keratoconus.

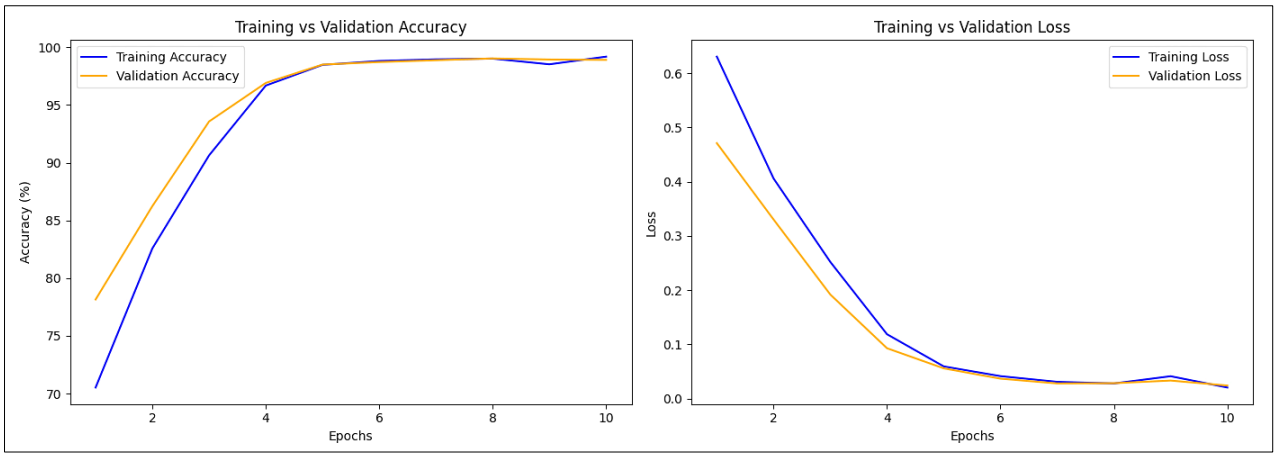


Figure 18. Accuracy and Loss graph of MobileNetV2+ViT

4-2-6- Accuracy and Loss Graph of ResNet50+ViT

Figure 19 shows the training and validation accuracy (left) and loss (right) of the ResNet50+ViT hybrid model, trained for 10 epochs. The accuracy graph represents rapid improvements in the first few epochs. The training and validation accuracy started at approximately 75% and 85%, respectively, then surpassed 97% for both in the 4th epoch, converging near 99% in the 6th epoch and remaining constant for the rest of the training. The loss plot reflects a great drop in training and validation loss in a few beginning epochs-the training loss dropped from about 0.5 to nearly 0.01 by the 6th epoch, and the validation loss did, too, in that trend. The closeness of the training and validation curves points to excellent generalization without evidence of overfitting. The ResNet50+ViT hybrid model shows in general fantastic overall performance, considering high accuracy and a minimum loss for keratoconus classification.

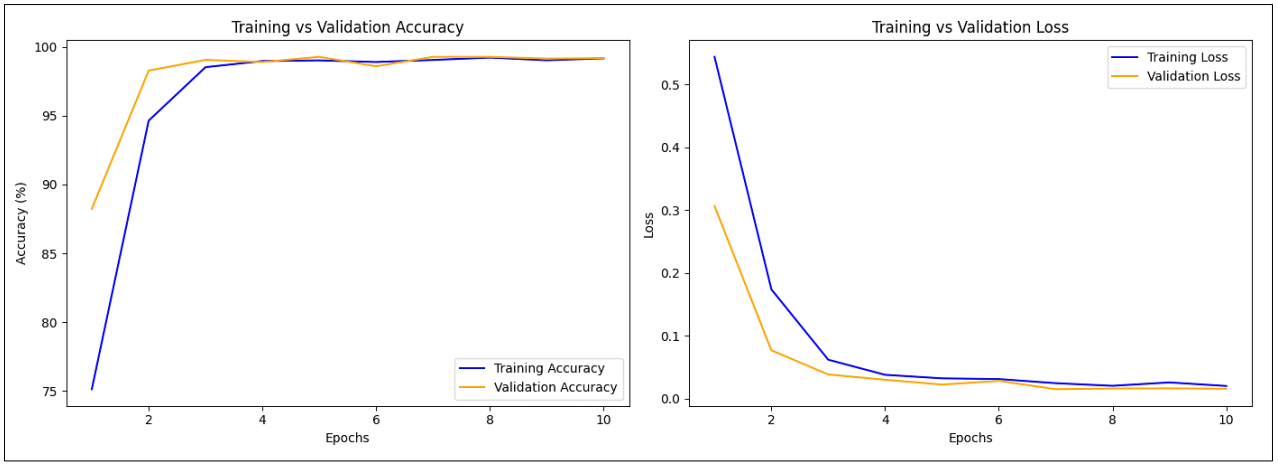


Figure 19. Accuracy and Loss graph of ResNet50+ViT

4-2-7- Accuracy and Loss Graph of VGG16+ViT

Figure 20 presents the training and validation accuracy (left) and loss (right) for the VGG16+ViT hybrid model over 10 epochs. The graph for accuracy increases very steeply during the first few epochs, while the training accuracy starts from approximately 70% and the validation accuracy from 80%. Both metrics increase very quickly, both reaching over 95% by the 3rd epoch, before leveling out at close to 99% around the 5th epoch, where they remain for the rest of training. This is reflected in a corresponding loss plot that has the training loss decrease from roughly 0.6 to near 0.01 by the 6th epoch. The validation loss follows suit, leveling out just above zero and barely blemishing a flat line to indicate very good generalization. The alignment of the training and validation curves underlines the efficiency and reliability of the model, confirming its capability for the accurate classification of keratoconus with minimal overfitting.

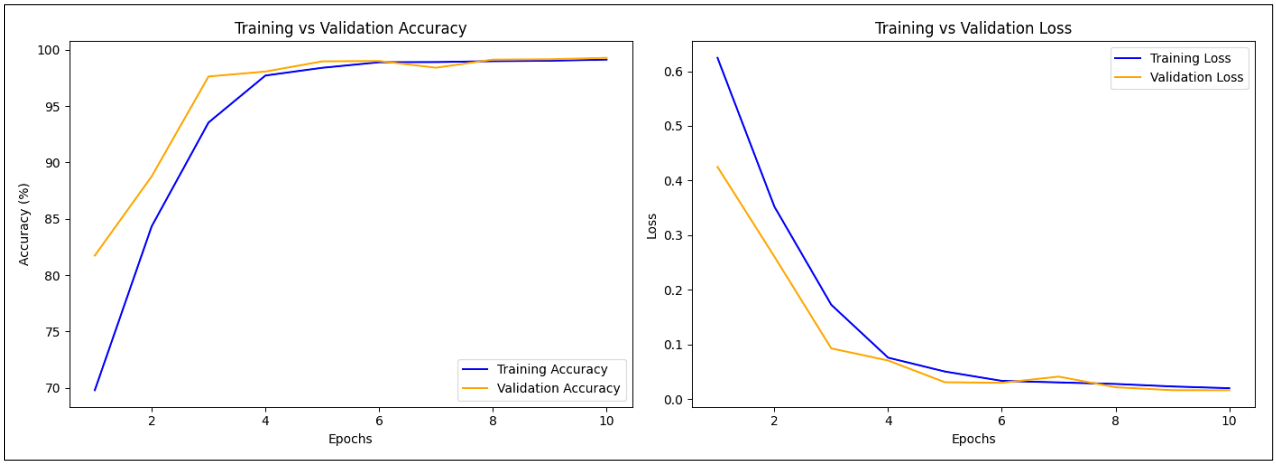


Figure 20. Accuracy and Loss graph of VGG16+ViT

4-2-8- Accuracy and Loss Graph of VGG19+ViT

Figure 21 shows, for 10 epochs, both the training and validation accuracy (left) and loss (right) of the VGG19+ViT hybrid model. From this accuracy plot, it can be seen that in the early epoch, there is a sudden increase; starting from about 70% in training accuracy while 80% in validation accuracy, then both started to surpass 95% by the 3rd epoch and maintained about 99% by the 5th epoch with the rest of the epochs. The loss plot indicates a sharp drop within the first few epochs-the training loss drops from approximately 0.6 to near 0.01 by the 6th epoch, while the validation loss follows this and stabilizes with minimal fluctuation. Good generalization is indicated by the close alignment of the training and validation curves; it learned quite well without overfitting. It should be noted that the VGG19+ViT hybrid model outperformed with overall high accuracy and low loss in classifying keratoconus.

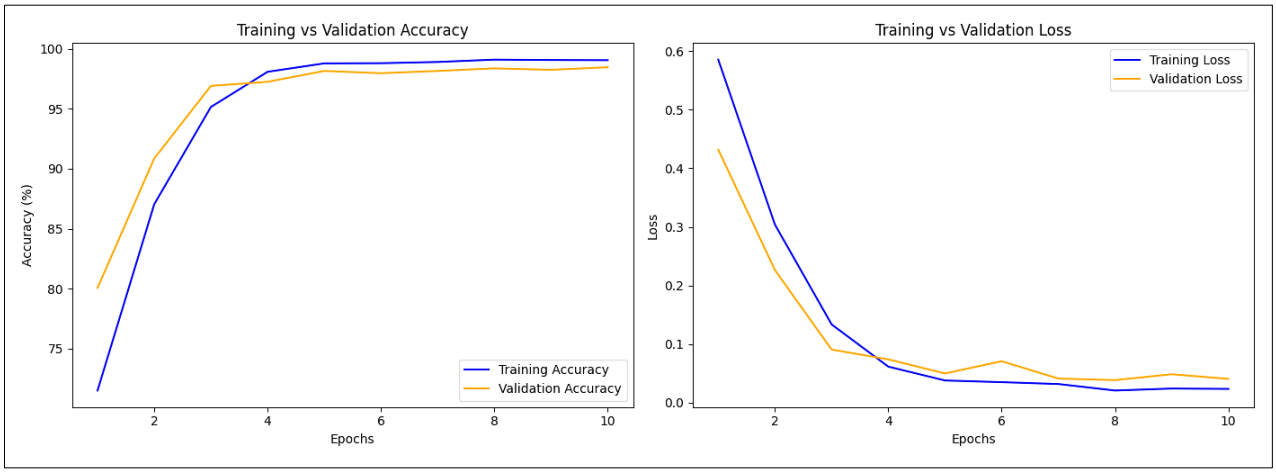


Figure 21. Accuracy and Loss graph of VGG19+ViT

4-3- Confusion Matrix

The confusion matrix provides a detailed evaluation of the model's performance by showing true positives, true negatives, false positives, and false negatives for each class. It helps identify misclassifications, understand class-specific weaknesses (e.g., distinguishing "Suspect" from "Normal"), and fine-tune the model, ensuring more balanced and reliable keratoconus disease classification. Confusion matrices of each model is discussed one by one.

4-3-1- Confusion Matrix of DenseNet121+ViT

The confusion matrix of the DenseNet121+ViT hybrid model shows very good classification performance for the three categories: Keratoconus, Normal, and Suspect. It is observed that most of the samples are correctly classified by the model, with 1,143 true positives for Keratoconus, 1,088 for Normal, and 950 for Suspect. Misclassifications are minimal, with only 3 Keratoconus samples being incorrectly labeled as Suspect, 6 Normal samples misclassified as

Suspect, and a slight confusion in the Suspect category, where 7 samples each are misclassified as Keratoconus and Normal. The overall evidence shows that the model is highly accurate and reliable, with very few false positives and negatives, hence robust for keratoconus classification (see Figure 22).

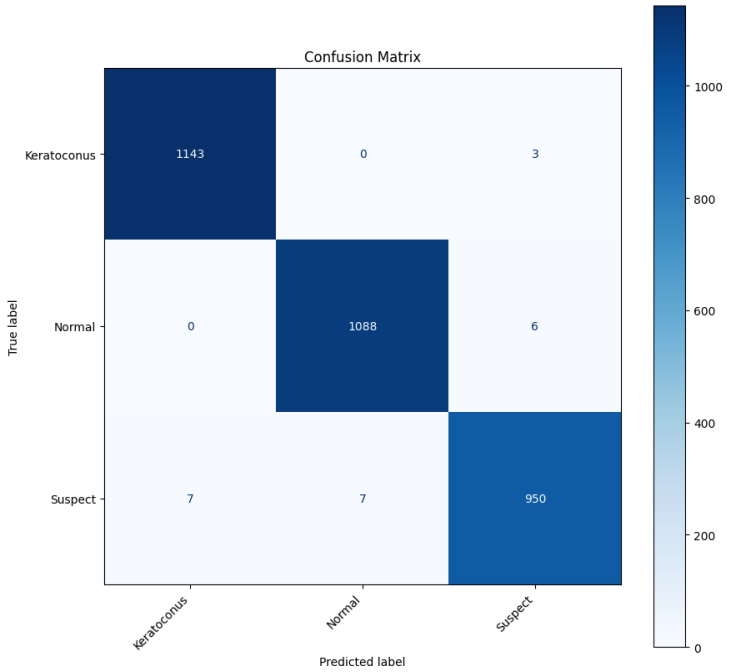


Figure 22. Confusion Matrix of DenseNet121+ViT

4-3-2- Confusion Matrix of EfficientNetB0+ViT

The confusion matrix of the EfficientNetB0+ViT hybrid model for Keratoconus, Normal, and Suspect classification is presented in Figure 23. This model classifies most samples correctly: 1,087 true positives for Keratoconus, 1,095 for Normal, and 985 for Suspect. Misclassifications are very few: only 2 keratoconus samples were classified as Suspect while 4 Normal ones as Keratoconus and 13 Normal ones as Suspect. Similarly, the Suspect class also shows slight confusion, with 8 samples being Keratoconus and 10 being Normal. Despite these minor errors, the model demonstrates excellent overall accuracy and robust performance in the classification of keratoconus, with the Keratoconus class showing near-perfect predictions.

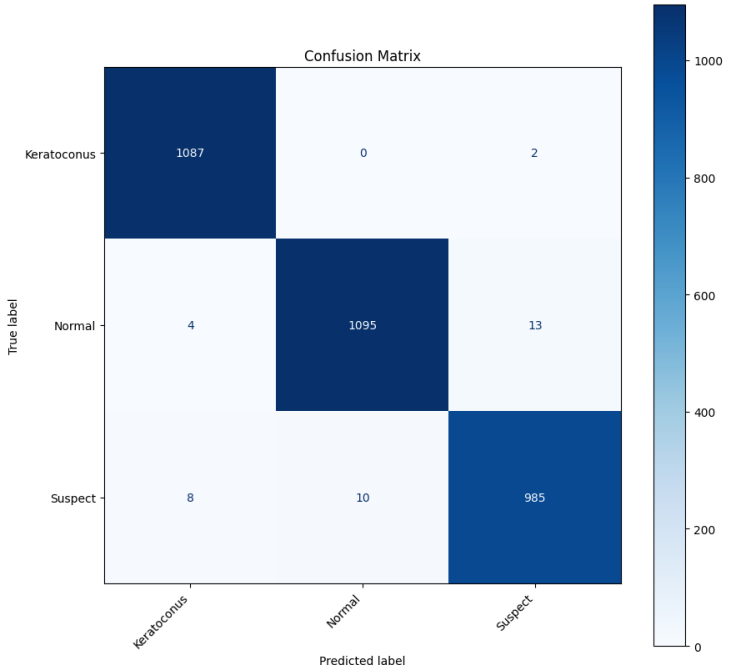


Figure 23. Confusion Matrix of EfficientNetB0+ViT

4-3-3- Confusion Matrix of InceptionResNetV2+ViT

The confusion matrix for the InceptionResNetV2+ViT hybrid model is presented in Figure 24, and it underlines its excellent performance in the classification among the three categories: Keratoconus, Normal, and Suspect. For the model, 1,042 true positives were classified correctly for Keratoconus, with only 2 being misclassified as Suspect, showing great accuracy in this critical category. In the case of the Normal class, 1,134 samples were correctly classified, with 19 misclassified as Suspect, showing a slight tendency toward false positives in this category. The Suspect class is also doing quite well: 991 are true positives, 2 have been misclassified as Keratoconus, and 14 as Normal. Low rates of misclassifications and the high alignment across the classes confirm robust model performance to yield a distinction between the classes with minimal errors.

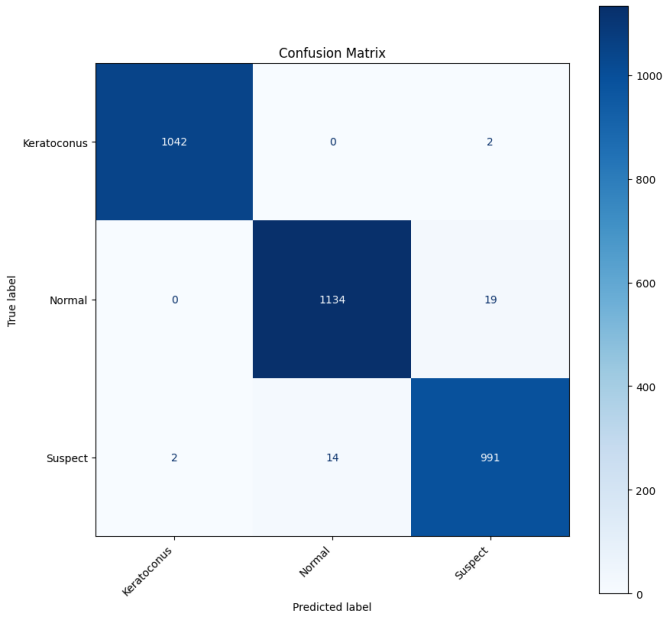


Figure 24. Confusion Matrix of InceptionResNetV2+ViT

4-3-4- Confusion Matrix of InceptionV3+ViT

The confusion matrix from the InceptionV3+ViT hybrid model presents, in detail, its strong performance in classifying the three categories: Keratoconus, Normal, and Suspect. It also classified 1,141 of the Keratoconus correctly, mislabeling only 4 as Suspect, indicating very high accuracy for this key class. Then, it gives 1,084 true positives for the Normal class, with a minimum of 7 being misclassified into the Suspect category. The Suspect class is well-identified, with 949 true positives, but it has slight confusion: 2 samples were misclassified as Keratoconus and 17 as Normal. In general, the model has excellent generalization and is very robust, since only a few errors appeared, and the performance is balanced for all classes, making the model highly reliable for keratoconus classification (see Figure 25).

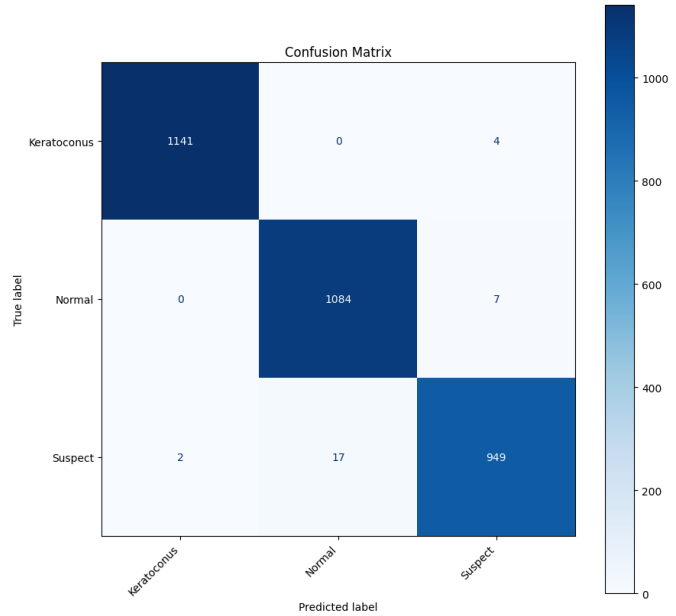


Figure 25. Confusion Matrix of InceptionV3+ViT

4-3-5- Confusion Matrix of MobileNetV2+ViT

The confusion matrix of the MobileNetV2+ViT hybrid model shows excellent classification performance within the three classes: Keratoconus, Normal, and Suspect. It classifies 1,120 Keratoconus samples as such, with only 6 being misclassified under Suspect. This is a very good value of precision within this class. For the class Normal, 1,093 are true positives, with a minimum misclassification-only 7 samples under Suspect. The Suspect class also has a very high degree of accuracy, with 956 true positives, but some confusion can be seen since 5 have been misclassified as Keratoconus and 17 as Normal. Overall, the model seems robust and shows strong generalization, effectively discriminating between classes with very few errors, confirming its reliability for the classification of keratoconus (see Figure 26).

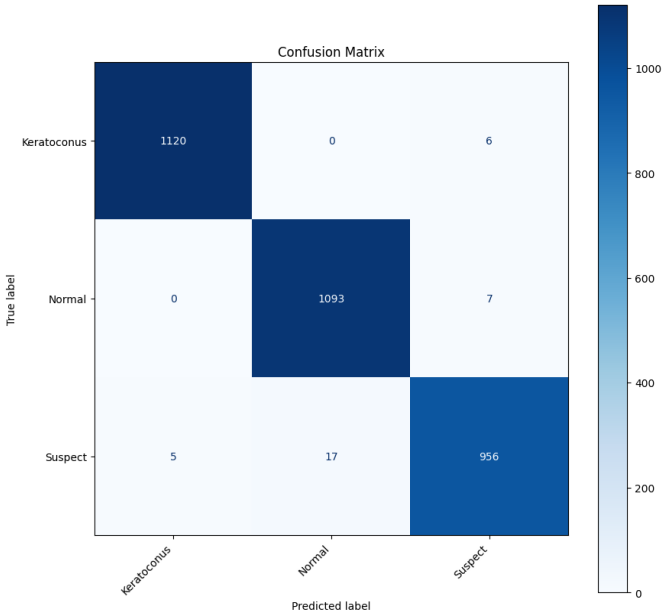


Figure 26. Confusion Matrix of MobileNetV2+ViT

4-3-6- Confusion Matrix of ResNet50+ViT

The confusion matrix of the ResNet50+ViT hybrid model shows great classification performance for the three classes: Keratoconus, Normal, and Suspect. It classified 1,086 samples correctly as Keratoconus, misclassifying only 1 as Suspect, which shows outstanding precision in this category. For the Normal class, it achieved 1,173 true positives, with just 4 samples misclassified as Suspect, showcasing near-perfect accuracy. The Suspect class is the best performing class with 919 true positives; however, the model seems a bit perplexed while considering 4 as Keratoconus and 17 as Normal. In this way, the ResNet50+ViT hybrid model obtained very good generalization and reliability regarding the distinction among classes, providing only a small number of mistakes. This constitutes a strong point in keratoconus classification (see Figure 27).

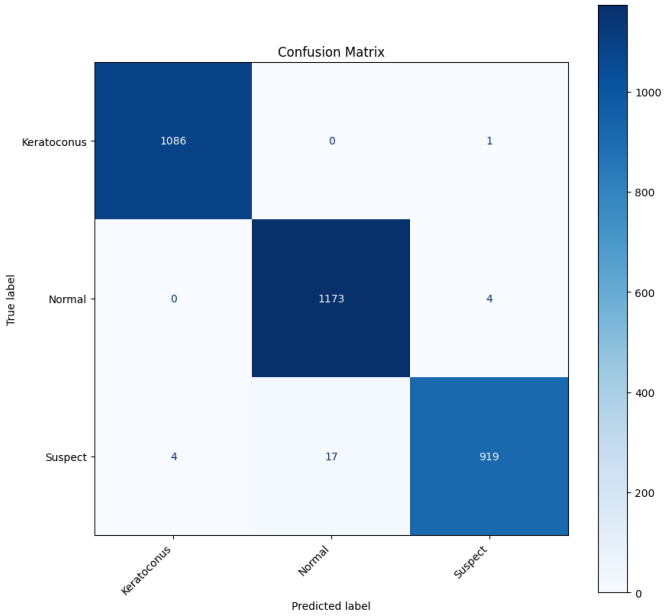


Figure 27. Confusion Matrix of ResNeet50+ViT

4-3-7- Confusion Matrix of VGG16+ViT

The confusion matrix of the VGG16+ViT hybrid model has shown excellent classification performance among the three categories: Keratoconus, Normal, and Suspect. For Keratoconus, the model correctly classifies 1,145 samples with only 4 misclassified as Suspect, showing very good precision for this class. In the case of the Normal class, it achieves 1,119 true positives with a minimum misclassification of 10 samples as Suspect. The Suspect class is well-represented, with 917 true positives; there are, however, minor confusions with the classes Keratoconus and Normal, misclassified with 2 and 7 samples, respectively. On the whole, the proposed hybrid model of VGG16+ViT demonstrates outstanding accuracy with excellent robustness, efficiently generalizing on all classes with limited errors, thus proving to be highly reliable for keratoconus classification (see Figure 28).

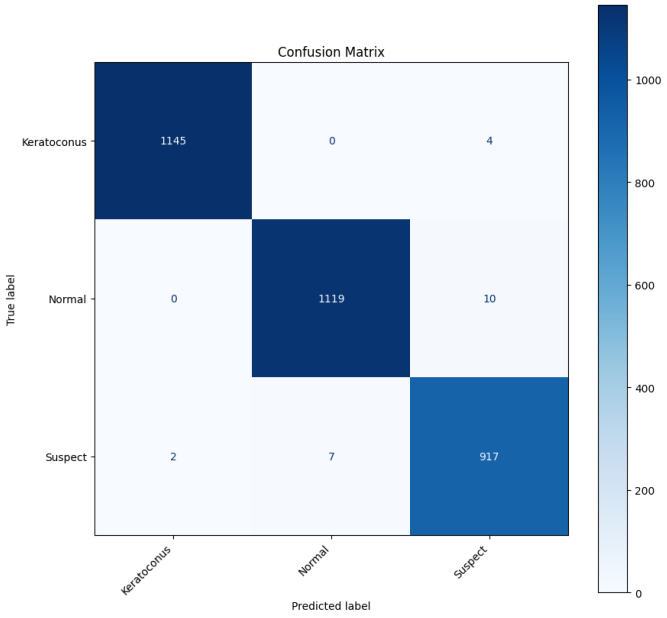


Figure 28. Confusion Matrix of VGG16+ViT

4-3-8- Confusion Matrix of VGG19+ViT

The confusion matrix of the proposed VGG19+ViT hybrid model shows very nice classification performance over the three classes: Keratoconus, Normal, and Suspect. It correctly identified 1,145 Keratoconus samples-only 4 samples were misinterpreted as Suspect-which implies high precision over this class. For the normal class, it achieved 1,119 true positives, where only 10 samples were misjudged as suspect. For the Suspect class, 917 are true positives, although there is slight confusion since 2 of them are misclassified into Keratoconus and 7 into Normal. Overall, the VGG19+ViT hybrid model generalizes well for all classes with only minor errors, hence being a potential model for keratoconus classification (see Figure 29).

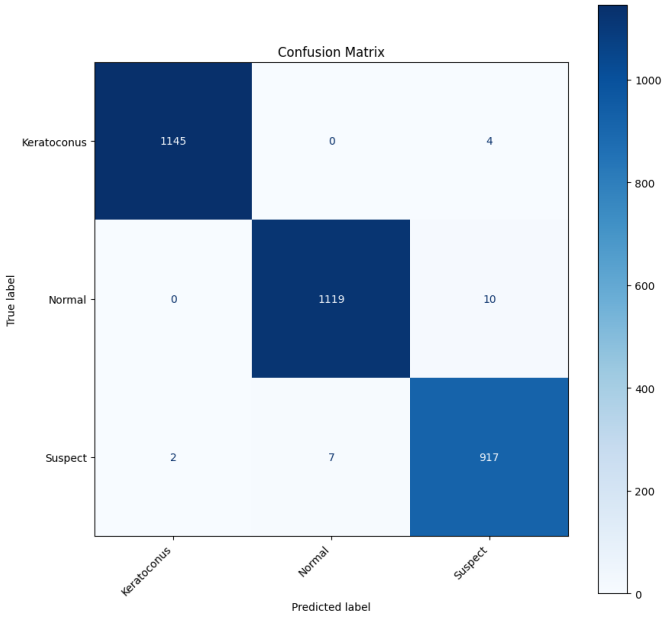


Figure 29. Confusion Matrix of VGG19+ViT

4-4- Analysis of the Overall Findings

Hybrid Models Significantly Outperform Standalone CNNs: One of the most prominent findings of this study is that combining pre-trained CNN models with the Vision Transformer (ViT) results in substantial performance gains across all key metrics—training accuracy, validation accuracy, testing accuracy, precision, recall, and F1-score. While standalone CNN models like DenseNet121, EfficientNetB0, InceptionResNetV2, and ResNet50 show moderate to high classification accuracy, their hybrid counterparts consistently demonstrate superior performance. This confirms that the integration of CNNs' strong local feature extraction capabilities with ViT's global attention mechanism provides a synergistic effect, enabling the models to capture both fine-grained details and long-range contextual dependencies in keratoconus classification. The observed accuracy improvements (often exceeding 30% over the standalone models) underscore the hybrid approach's potential in overcoming the limitations inherent in individual CNN architectures.

The hybrid CNN-ViT model naturally introduces additional computational complexity compared to standalone CNNs or ViTs due to the sequential processing of features—first through the CNN feature extractor and subsequently through the ViT classifier. However, to mitigate excessive computational demands, pretrained CNN models were utilized for feature extraction, effectively reducing the need for training these layers from scratch. Additionally, the ViT component, being relatively shallow, ensures that the overall model remains computationally manageable. While the hybrid model's training time is moderately longer than standalone CNNs, it achieves significant performance gains, particularly in generalization and accuracy. The experiments showed that although training time increased by approximately 20-30%, the accuracy improvement of 10-33% justified this trade-off. Moreover, the batch size and number of epochs were carefully selected to balance computational load and model performance.

DenseNet121+ViT and ResNet50+ViT Show the Most Dramatic Improvements: Among the various hybrid configurations evaluated, DenseNet121+ViT and ResNet50+ViT exhibit the most significant performance boosts. DenseNet121, when used alone, achieves only 69% testing accuracy, but when integrated with ViT, the accuracy jumps to 99.28%. Similarly, ResNet50's standalone performance of 66% testing accuracy improves dramatically to 99.18% when hybridized with ViT. This marked improvement can be attributed to the dense connectivity and residual learning properties of DenseNet121 and ResNet50, which, when enhanced by ViT's capacity to model long-range dependencies and relationships, results in robust classification. These findings highlight that the global context awareness offered by ViT effectively complements CNNs' hierarchical feature learning, particularly in architectures that might otherwise struggle with deeper or more complex feature relationships.

Exceptional Performance in Keratoconus Classification: Across all hybrid models, the classification of the Keratoconus class achieves perfect or near-perfect precision, recall, and F1-scores (all around 1.0). This is a critical outcome, as keratoconus is the primary focus of this research. Models like DenseNet121+ViT, ResNet50+ViT, InceptionV3+ViT, and VGG16+ViT achieve flawless classification in this category, with no false positives or false negatives. Such consistently high performance demonstrates that the hybrid models effectively learn and generalize the distinct features associated with Keratoconus, ensuring minimal diagnostic errors. This reliability is especially vital in real-world clinical applications, where accurate disease detection directly impacts patient outcomes.

Slight Performance Variability in Normal and Suspect Classes: While keratoconus classification performance is outstanding, the Normal and Suspect classes exhibit minor fluctuations in precision and recall across some models. Specifically, models like EfficientNetB0+ViT and InceptionResNetV2+ViT show slight drops in recall for the Normal class (around 0.98), and a similar trend is seen for the Suspect class. These misclassifications can be attributed to the inherent difficulty in distinguishing Suspect cases, which may share overlapping characteristics with both Normal and Keratoconus categories. Nonetheless, these models still maintain high precision, indicating their ability to minimize false positives, though further work may be necessary to enhance sensitivity towards these borderline cases.

Robust Generalization Without Overfitting: The training and validation accuracy and loss curves across all hybrid models demonstrate excellent convergence behavior. Both training and validation accuracies reach near 99% within the first few epochs, while the loss stabilizes at a minimal value, showing no signs of overfitting or underfitting. The close alignment of training and validation curves confirms the models' strong generalization capabilities, meaning that the models perform equally well on unseen data as they do on training data. This robustness is crucial for ensuring reliable performance in diverse clinical settings.

Complementary Strengths of CNN and ViT: The overall findings emphasize the complementary strengths of CNN architecture and the Vision Transformer. While CNNs excel at local feature extraction and hierarchical pattern recognition, they may lack the ability to capture complex, long-range dependencies across an image. ViT compensates for this by applying self-attention mechanisms, providing a broader contextual understanding. This combination creates a powerful hybrid model capable of simultaneously leveraging detailed local features and global contextual information, which is particularly beneficial in handling complex, high-variance medical datasets like keratoconus images.

Potential for Real-World Clinical Deployment: Given the consistently high accuracy, precision, and generalization capability demonstrated by the hybrid models, particularly in Keratoconus detection, these models are highly suitable

for real-world clinical applications. Their ability to deliver precise and reliable classifications with minimal misclassification rates suggests strong potential for integration into computer-aided diagnostic systems, assisting ophthalmologists and clinicians in making faster, more accurate decisions.

4-5- Comparison with Other Studies

Table 3 provides a comprehensive comparison of various recent models applied to keratoconus classification, emphasizing the diversity of techniques, dataset sizes, and achieved accuracies. When critically evaluating the models in relation to the proposed Vision Transformer Embedded Feature Fusion Model, several performance trends and strengths of the proposed approach become evident.

To begin with, conventional machine learning and deep learning models such as CNNs, SVMs, Random Forests, Naive Bayes, and k-NN have been employed in earlier studies with varying degrees of success. For instance, Hashim and Mazinani (2025) used multiple models on a dataset of 400 images across two classes. Among these, CNN achieved the highest accuracy of 99%, which demonstrates the power of deep learning over traditional ML techniques. However, this performance was achieved on a relatively small binary-class dataset, limiting the generalizability of the model to more complex, real-world scenarios involving multiple classes and larger datasets.

Similarly, Askarian et al. (2024) and Wan et al. (2025) applied traditional classifiers and boosting techniques on small datasets (100 and 95 images respectively) and reported high accuracies of 97% and 98%. Despite these promising results, the limited dataset sizes introduce a risk of overfitting and poor scalability. Furthermore, these models did not incorporate advanced deep learning or attention-based mechanisms that can capture the subtle patterns inherent in keratoconus progression.

Table 3. Comparison of the performances of the proposed models against a few recent published works

Author(s)	Dataset size	Model type	Accuracy
Hashim & Mazinani (2025) [46]	400 images (2 classes)	CNN, Naive Bayes, KNN, AdaBoost, Decision Trees	CNN - 99%
Yaraghi & Khatibi, (2024) [50]	92 images (2 classes)	Fusion of pretrained models (InceptionRestNetV2, VGG16, EfficientNetB0) and Vision Transformer (ViT)	Hybrid model (pretrained models + ViT - 96.20%)
Hartmann et al. (2024) [49]	570 images (2 classes)	MLP, Three convolutional layers (CNN), MobileNet + MLP	MobileNet + MLP - 83%
Wan et al. (2025) [45]	95 images (2 classes)	LASSO, XGBoost, Random Forest (RF) for feature selection leading to a nomogram	XGBoost – 98%
Ismael (2025) [47]	size not specified	Hybrid Transformer architecture (Vision Transformer and Transformer Encoder), SVM, Random Forests, CNN	Hybrid Transformer architecture -98.48%
Askarian et al. (2024) [48]	100 images (2 classes)	Linear SVM, Gaussian SVM, CNNs, k-NN	Gaussian SVM - **97%**
Shi et al. (2024) [51]	573 images (3 classes)	Feature Vector Aggregation Network (FVAnet) using EfficientNet, ResNet-50, CNN fusion, voting	FVAnet - **78.7%**
Al-Timemy et al. (2021) [44]	4844 corneal images (2 classes)	EfficientNet-B0	97.70%
Al-Timemy et al. (2023) [7]	1371 eyes (3 maps)	Xception and InceptionResNetV2	99%
Proposed Model	4011 images (3 classes)	Feature Fusion model (Pretrained Models +ViT)	**DenseNet121+ViT and VGG16+ViT – 99.28%**

Moving to more recent advancements, hybrid models integrating pre-trained CNNs with Transformer-based architectures have been explored. Yaraghi & Khatibi [50], for example, proposed a fusion of InceptionResNetV2, VGG16, EfficientNetB0, and Vision Transformer (ViT) models on a small dataset of 92 images, achieving 96.20% accuracy. Likewise, Ismael [47] applied a hybrid Transformer model combining ViT, Transformer Encoders, and CNN, reporting a robust accuracy of 98.48%. These models show the effectiveness of combining CNN and Transformer features, but their limitations lie in the small dataset sizes and binary classification focus, which restrict their applicability in multi-class real-world clinical settings.

Further, Hartmann et al. [49] utilized a MobileNet + MLP combination, achieving only 83% accuracy on 570 images. This comparatively lower performance highlights the limitations of lighter architectures like MobileNet, especially when handling more nuanced corneal data without deeper feature extraction or global context modeling.

The work of Al-Timemy et al. [7, 44] stands out as one of the most relevant benchmarks. In Al-Timemy et al. [44], EfficientNet-b0 achieved 97.70% accuracy on a sizable dataset of 4,844 images. In Al-Timemy et al. [7], using a deep feature fusion approach combining Xception and InceptionResNetV2 across three corneal maps, they achieved an impressive 99% accuracy. Although highly competitive, these studies primarily focused on CNN architectures, which are inherently limited in capturing long-range dependencies and contextual relationships in the data.

In contrast, the Proposed Model addresses several limitations observed in prior work. It is trained on a relatively large dataset comprising 4,011 images across three classes, providing a more robust evaluation and greater generalizability. By integrating pre-trained CNN models (DenseNet121 and VGG16) with Vision Transformers, the model benefits from both local feature extraction and global contextual learning. The feature fusion approach further enhances the model's capability to harness complementary information from multiple sources. As a result, the proposed model achieves the highest reported accuracy of 99.28%, outperforming other studies, especially in the more challenging multi-class classification scenario.

Moreover, the proposed model's design emphasizes scalability and adaptability to real-world clinical applications. While earlier models demonstrate competitive performance on small datasets and binary classifications, they fall short in addressing the complexity of multi-class keratoconus staging. The proposed model's superior performance, both in terms of accuracy and dataset size, suggests it is better equipped to meet the demands of modern ophthalmic diagnostics.

Finally, it can be concluded that, while previous models have contributed valuable insights into keratoconus classification, the proposed Vision Transformer Embedded Feature Fusion Model sets a new benchmark by achieving high accuracy in a multi-class, large dataset environment, combining the strengths of CNNs and Transformers for comprehensive feature representation. This makes it a promising solution for clinical implementation, particularly for early detection and staging of keratoconus.

5- Conclusion and Future Scopes

This study presented a Vision Transformer Embedded Feature Fusion Model that effectively combines the local feature extraction capabilities of pretrained Convolutional Neural Networks (CNNs) with the global contextual learning strengths of Vision Transformers (ViTs) for the automated classification of keratoconus (KCN) into Normal, Suspect, and Keratoconus categories. The hybrid architecture demonstrated superior performance in terms of classification accuracy, benefiting from the complementary strengths of CNNs and ViTs. Comprehensive preprocessing steps, including quality filtering, image normalization, and data augmentation, contributed to the robustness and reliability of the model, while rigorous evaluation on publicly available datasets affirmed its effectiveness.

A key consideration moving forward is the translation of this model from experimental settings to real-world clinical environments. One potential direction involves embedding the developed model into existing diagnostic software systems used by ophthalmologists. Integrating the model within clinical workflows would allow ophthalmologists to utilize automated, real-time keratoconus classification alongside traditional diagnostic procedures. Such integration could streamline the diagnostic process, reduce clinician workload, and minimize inter-observer variability, ultimately facilitating early and accurate identification of keratoconus cases. Furthermore, real-world deployment would require adherence to regulatory and ethical guidelines, ensuring that patient data privacy and system reliability are maintained at all times.

In addition to software integration, another avenue worth exploring is the inclusion of multimodal ophthalmic imaging data to further enhance the model's diagnostic capability. Currently, the model primarily utilizes corneal topography data, which effectively captures surface curvature and shape. However, incorporating additional imaging modalities such as Optical Coherence Tomography (OCT) scans or corneal tomography could provide more detailed information about the corneal structure and biomechanical properties. OCT imaging, in particular, offers high-resolution cross-sectional images of the corneal layers, which could reveal subtle structural changes indicative of early-stage or subclinical keratoconus. Combining features extracted from topography, OCT, and possibly biomechanical assessments could improve the model's ability to distinguish between borderline cases and normal variations, leading to higher diagnostic accuracy.

Moreover, expanding the dataset to include more diverse patient populations and imaging conditions is essential to ensure the generalizability and reliability of the model across different clinical settings. This may involve conducting multi-center collaborations and acquiring datasets from various demographics and imaging systems. The inclusion of diverse data will not only improve the robustness of the model but also allow for broader applicability across global healthcare environments.

Interpretability remains another critical focus for future development. Deep learning models, including ViTs, are often criticized for their "black-box" nature, which can hinder clinical trust. Enhancing the interpretability of the model's decisions—through techniques such as visualization of attention maps or explainable AI frameworks—will allow clinicians to understand the rationale behind each classification. Providing clear, human-interpretable explanations will be vital in building confidence among healthcare professionals and encouraging the adoption of AI-based diagnostic tools in everyday practice.

Lastly, thorough clinical validation is necessary before integrating the model into routine ophthalmic care. Large-scale, prospective studies involving ophthalmologists and patients in real clinical settings can assess the model's performance, usability, and impact on diagnostic outcomes. Such validations will be crucial for regulatory approval and eventual clinical deployment.

6- Declarations

6-1- Author Contributions

Conceptualization, M.S.R.; methodology, N.A.; software, M.F.I.; validation, M.M.R.; formal analysis, M.I.K.J.; investigation, A.T.; resources, S.A.; data curation, D.M.T.; writing—original draft preparation, N.A.; writing—review and editing, M.S.R.; visualization, S.J.; supervision, M.S.R.; project administration, N.A.; funding acquisition, M.S.R. All authors have read and agreed to the published version of the manuscript.

6-2- Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6-3- Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6-4- Acknowledgements

We are grateful to the Isphani Islamia Eye Institute and Hospital for guiding us throughout research. Besides, we appreciate the work of our earlier researchers, who sparked our interest in this topic.

6-5- Institutional Review Board Statement

This study was approved by the Faculty of Science and Information Technology at Daffodil International University, Bangladesh.

6-6- Informed Consent Statement

An explanatory statement was given to the participants to read in which their rights and risks were outlined. After being satisfied with the statement, they signed the informed consent form.

6-7- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

7- References

[1] Namba, H., Maeda, N., Utsunomiya, H., Kaneko, Y., Ishizawa, K., Ueno, Y., & Nishitsuka, K. (2025). Prevalence of keratoconus and keratoconus suspect, and their characteristics on corneal tomography in a population-based study. PLoS ONE, 20(1 January), 308892. doi:10.1371/journal.pone.0308892.

[2] Chen, X., Zhao, J., Iselin, K. C., Borroni, D., Romano, D., Gokul, A., McGhee, C. N. J., Zhao, Y., Sedaghat, M. R., Momeni-Moghaddam, H., Ziaei, M., Kaye, S., Romano, V., & Zheng, Y. (2021). Keratoconus detection of changes using deep learning of colour-coded maps. BMJ Open Ophthalmology, 6(1), 824. doi:10.1136/bmjophth-2021-000824.

[3] Li, S., Huo, Y., Xie, R., Han, Y., Zou, H., & Wang, Y. (2025). Enhancing early detection of keratoconus suspects using interocular corneal tomography asymmetry. International Ophthalmology, 45(1), 55. doi:10.1007/s10792-025-03423-7.

[4] Kuo, B. I., Chang, W. Y., Liao, T. S., Liu, F. Y., Liu, H. Y., Chu, H. S., Chen, W. L., Hu, F. R., Yen, J. Y., & Wang, I. J. (2020). Keratoconus screening based on deep learning approach of corneal topography. Translational Vision Science and Technology, 9(2), 1–11. doi:10.1167/tvst.9.2.53.

[5] Kamiya, K., Ayatsuka, Y., Kato, Y., Shoji, N., Miyai, T., Ishii, H., Mori, Y., & Miyata, K. (2021). Prediction of keratoconus progression using deep learning of anterior segment optical coherence tomography maps. Annals of Translational Medicine, 9(16), 1287–1287. doi:10.21037/atm-21-1772.

[6] Muhsin, Z., Qahwaji, R., Ghafir, I., AlShawabkeh, M. A., Al Bdour, M., AlRyalat, S. A., & Al-Taee, M. Two-Stage Ensemble Learning Framework for Automated Classification of Keratoconus Severity. SSRN Electronic Journal, 1-17. doi:10.2139/ssrn.5096562.

[7] Al-Timemy, A. H., Alzubaidi, L., Mosa, Z. M., Abdelmotaal, H., Ghaeb, N. H., Lavric, A., Hazarbassanov, R. M., Takahashi, H., Gu, Y., & Yousefi, S. (2023). A Deep Feature Fusion of Improved Suspected Keratoconus Detection with Deep Learning. Diagnostics, 13(10), 1689–1689,. doi:10.3390/diagnostics13101689.

[8] Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y. W., & Wu, J. (2020). UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2020-May, 1055–1059. doi:10.1109/ICASSP40776.2020.9053405.

[9] Gu, H., Guo, Y., Gu, L., Wei, A., Xie, S., Ye, Z., Xu, J., Zhou, X., Lu, Y., Liu, X., & Hong, J. (2020). Deep learning for identifying corneal diseases from ocular surface slit-lamp photographs. Scientific Reports, 10(1), 17851. doi:10.1038/s41598-020-75027-3.

[10] Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-January, 5987–5995. doi:10.1109/CVPR.2017.634.

[11] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. 36th International Conference on Machine Learning, ICML 2019, 2019-June, 10691–10700.

[12] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, 770–778. doi:10.1109/CVPR.2016.90.

[13] Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, inception-ResNet and the impact of residual connections on learning. 31st AAAI Conference on Artificial Intelligence, AAAI 2017, 31(1), 4278–4284. doi:10.1609/aaai.v31i1.11231.

[14] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, 2818–2826. doi:10.1109/CVPR.2016.308.

[15] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 4510–4520. doi:10.1109/CVPR.2018.00474.

[16] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings, 1-14. doi:10.48550/arXiv.1409.1556

[17] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, 248–255. doi:10.1109/CVPR.2009.5206848.

[18] Deheyab, A. O. A., Alwan, M. H., Rezzaqe, I. K. A., Mahmood, O. A., Hammadi, Y. I., Kareem, A. N., & Ibrahim, M. (2022). An Overview of Challenges in Medical Image Processing. ACM International Conference Proceeding Series, 511–516. doi:10.1145/3584202.3584278.

[19] Al Bdour, M., Sabbagh, H. M., & Jammal, H. M. (2024). Multi-modal imaging for the detection of early keratoconus: a narrative review. Eye and Vision, 11(1), 18. doi:10.1186/s40662-024-00386-1.

[20] Al-Timemy, A. H., Mosa, Z. M., Alyasseri, Z., Lavric, A., Lui, M. M., Hazarbassanov, R. M., & Yousefi, S. (2021). A Hybrid Deep Learning Construct for Detecting Keratoconus From Corneal Maps. Translational Vision Science and Technology, 10(14), 16. doi:10.1167/tvst.10.14.16.

[21] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021 - 9th International Conference on Learning Representations, 1-22. doi:10.48550/arXiv.2010.11929.

[22] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. Proceedings of Machine Learning Research, 139, 10347–10357. doi:10.48550/arXiv.2012.12877.

[23] He, K., Gan, C., Li, Z., Rekik, I., Yin, Z., Ji, W., Gao, Y., Wang, Q., Zhang, J., & Shen, D. (2023). Transformers in medical image analysis. Intelligent Medicine, 3(1), 59–78. doi:10.1016/j.imed.2022.07.002.

[24] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2323. doi:10.1109/5.726791.

[25] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345–1359. doi:10.1109/TKDE.2009.191.

[26] Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 8689 LNCS, Issue PART 1, 818–833. doi:10.1007/978-3-319-10590-1_53.

[27] Xu, Y., Quan, R., Xu, W., Huang, Y., Chen, X., & Liu, F. (2024). Advances in Medical Image Segmentation: A Comprehensive Review of Traditional, Deep Learning and Hybrid Approaches. Bioengineering, 11(10), 1034. doi:10.3390/bioengineering11101034.

[28] Islam, S., Elmekki, H., Elsebai, A., Bentahar, J., Drawel, N., Rjoub, G., & Pedrycz, W. (2024). A comprehensive survey on applications of transformers for deep learning tasks. Expert Systems with Applications, 241, 122666–122666,. doi:10.1016/j.eswa.2023.122666.

[29] Al-hammuri, K., Gebali, F., Kanan, A., & Chelvan, I. T. (2023). Vision transformer architecture and applications in digital health: a tutorial and survey. Visual Computing for Industry, Biomedicine, and Art, 6(1), 14. doi:10.1186/s42492-023-00140-9.

[30] Rayed, M. E., Islam, S. M. S., Niha, S. I., Jim, J. R., Kabir, M. M., & Mridha, M. F. (2024). Deep learning for medical image segmentation: State-of-the-art advancements and challenges. Informatics in Medicine Unlocked, 47, 101504–101504. doi:10.1016/j.imu.2024.101504.

[31] Pinto-Coelho, L. (2023). How Artificial Intelligence Is Shaping Medical Imaging Technology: A Survey of Innovations and Applications. Bioengineering, 10(12), 1435. doi:10.3390/bioengineering10121435.

[32] Srinivasan, S., Francis, D., Mathivanan, S. K., Rajadurai, H., Shivahare, B. D., & Shah, M. A. (2024). A hybrid deep CNN model for brain tumor image multi-classification. BMC Medical Imaging, 24(1), 21. doi:10.1186/s12880-024-01195-7.

[33] Tamilselvi, S., Suchetha, M., & Raman, R. (2025). Leveraging ResNet50 with Swin Attention for Accurate Detection of OCT Biomarkers Using Fundus Images. IEEE Access, 3544332. doi:10.1109/ACCESS.2025.3544332.

[34] Azad, R., Jia, Y., Aghdam, E. K., Cohen-Adad, J., & Merhof, D. (2023). Enhancing Medical Image Segmentation with TransCeption: A Multi-Scale Feature Fusion Approach. arxiv:2301.10847, 1-11. doi:10.48550/arxiv.2301.10847.

[35] Gürsoy, E., & Kaya, Y. (2025). Multi-source deep feature fusion for medical image analysis. Multidimensional Systems and Signal Processing, 36(1), 1-20. doi:10.1007/s11045-024-00897-z.

[36] Patel, S., Patel, R., Ganatra, N., & Patel, A. (2022). Spatial Feature Fusion for Biomedical Image Classification based on Ensemble Deep CNN and Transfer Learning. International Journal of Advanced Computer Science and Applications, 13(5), 153–159. doi:10.14569/IJACSA.2022.0130519.

[37] Öksüz, C., Urhan, O., & Güllü, M. K. (2022). Brain tumor classification using the fused features extracted from expanded tumor region. Biomedical Signal Processing and Control, 72, 103356. doi:10.1016/j.bspc.2021.103356.

[38] Matsoukas, C., Haslum, J. F., Söderberg, M., & Smith, K. (2021). Is it Time to Replace CNNs with Transformers for Medical Images? arXiv:2108.09038, 1-6. doi:10.48550/arXiv.2108.09038.

[39] Yamanakkanavar, N., & Lee, B. (2022). MF2-Net: A multipath feature fusion network for medical image segmentation. Engineering Applications of Artificial Intelligence, 114, 105004. doi:10.1016/j.engappai.2022.105004.

[40] Zhou, T., Cheng, Q. R., Lu, H. L., Li, Q., Zhang, X. X., & Qiu, S. (2023). Deep learning methods for medical image fusion: A review. Computers in Biology and Medicine, 160, 106959. doi:10.1016/j.compbiomed.2023.106959.

[41] Khan, S. A., Khan, M. A., Song, O.-Y., & Nazir, M. (2020). Medical Imaging Fusion Techniques: A Survey Benchmark Analysis, Open Challenges and Recommendations. Journal of Medical Imaging and Health Informatics, 10(11), 2523–2531. doi:10.1166/jmihi.2020.3222.

[42] Lavric, A., & Valentin, P. (2019). KeratoDetect: Keratoconus Detection Algorithm Using Convolutional Neural Networks. Computational Intelligence and Neuroscience, 8162567, 1–9. doi:10.1155/2019/8162567.

[43] Yousefi, S., Yousefi, E., Takahashi, H., Hayashi, T., Tampo, H., Inoda, S., Arai, Y., & Asbell, P. (2019). Keratoconus severity identification using unsupervised machine learning. PLoS ONE, 13(11), 205998. doi:10.1371/journal.pone.0205998.

[44] Al-Timemy, A. H., Ghaeb, N. H., Mosa, Z. M., & Escudero, J. (2022). Deep Transfer Learning for Improved Detection of Keratoconus using Corneal Topographic Maps. Cognitive Computation, 14(5), 1627–1642. doi:10.1007/s12559-021-09880-3.

[45] Wan, Q., Wang, Q., Wei, R., Tang, J., Yin, H., Deng, Y. P., & Ma, K. (2025). Machine learning-based progress prediction in accelerated cross-linking for Keratoconus. Graefe's Archive for Clinical and Experimental Ophthalmology, 1-15. doi:10.1007/s00417-025-06792-y.

[46] Hashim, A. A., & Mazinani, M. (2025). Detection of Keratoconus Disease Depending on Corneal Topography Using Deep Learning. Kufa Journal of Engineering, 16(1), 463–478. doi:10.30572/2018/KJE/160125.

[47] Ismael, O. (2025). Enhancing keratoconus detection with transformer technology and multi-source integration. Artificial Intelligence Review, 58(1), 1-31. doi:10.1007/s10462-024-11016-6.

[48] Askarian, B., Askarian, A., Tabei, F., & Chong, J. W. (2025). An IoT-Enabled mHealth Sensing Approach for Remote Detection of Keratoconus Using Smartphone Technology. Sensors, 25(5), 1316. doi:10.3390/s25051316.

[49] Hartmann, L. M., Langhans, D. S., Eggarter, V., Freisenich, T. J., Hillenmayer, A., König, S. F., Vounotrypidis, E., Wolf, A., & Wertheimer, C. M. (2024). Keratoconus Progression Determined at the First Visit: A Deep Learning Approach With Fusion of Imaging and Numerical Clinical Data. Translational Vision Science and Technology, 13(5), 7–7,. doi:10.1167/tvst.13.5.7.

[50] Yaraghi, S., & Khatibi, T. (2024). Keratoconus disease classification with multimodel fusion and vision transformer: a pretrained model approach. BMJ Open Ophthalmology, 9(1), 1589. doi:10.1136/bmjophth-2023-001589.

[51] Shi, K., Dou, N., Sun, S., Li, M., Li, P., Xu, L., ... & Mi, S. Feature Vector Aggregation Network: A New Paradigm for Keratoconus Detection. SSRN Electronic Journal, 1-14. doi:10.2139/ssrn.4713088.

[52] Ji, Q., Huang, J., He, W., & Sun, Y. (2019). Optimized deep convolutional neural networks for identification of macular diseases from optical coherence tomography images. Algorithms, 12(3), 51. doi:10.3390/a12030051.

[53] Isaksson, L. J., Pepa, M., Summers, P., Zaffaroni, M., Vincini, M. G., Corrao, G., Mazzola, G. C., Rotondi, M., Lo Presti, G., Raimondi, S., Gandini, S., Volpe, S., Haron, Z., Alessi, S., Pricolo, P., Mistretta, F. A., Luzzago, S., Cattani, F., Musi, G., … Jereczek-Fossa, B. A. (2023). Comparison of automated segmentation techniques for magnetic resonance images of the prostate. BMC Medical Imaging, 23(1), 32. doi:10.1186/s12880-023-00974-y.

[54] Peng, C., Liu, Y., Yuan, X., & Chen, Q. (2022). Research of image recognition method based on enhanced inception-ResNet-V2. Multimedia Tools and Applications, 81(24), 34345–34365. doi:10.1007/s11042-022-12387-0.

[55] Iparraguirre-Villanueva, O., Guevara-Ponce, V., Paredes, O. R., Sierra-Liñan, F., Zapata-Paulini, J., & Cabanillas-Carbonell, M. (2022). Convolutional Neural Networks with Transfer Learning for Pneumonia Detection. International Journal of Advanced Computer Science and Applications, 13(9), 544–551. doi:10.14569/IJACSA.2022.0130963.

[56] Al-Rammahi, A. H. I. (2022). Face mask recognition system using MobileNetV2 with optimization function. Applied Artificial Intelligence, 36(1), 2145638. doi:10.1080/08839514.2022.2145638.

[57] Mukherjee, S. (2022). The annotated RESNet-50 - towards data science. Towards Data Science, California, United States. Available online: https://towardsdatascience.com/the-annotated-resnet-50-a6c536034758 (accessed on March 2025).

[58] Aluka, M., Reddy, V. P., & Ganesan, S. (2022). Comparative Analysis of CNN Regularisation and Augmentation Techniques with Ten Layer Deep Learning Model to Detect Lung Cancer. International Journal on Recent and Innovation Trends in Computing and Communication, 10(11), 33–39. doi:10.17762/ijritcc.v10i11.5777.

[59] Zheng, Y., Yang, C., & Merkulov, A. (2018). Breast cancer screening using convolutional neural network and follow-up digital mammography. Computational Imaging III, 4. doi:10.1117/12.2304564.