Improving CADx System Performance for Skin Disease Detection using Ensemble Machine Learning Models

Abu Asaduzzaman¹, Christian C. Thompson¹, and Md J. Uddin¹

¹Affiliation not available

May 02, 2024

Abstract

Conventional computer-aided diagnosis (CADx) systems play a crucial role in assisting medical professionals with the detection of skin diseases. However, these systems often involve manual, time-consuming, and error-prone processes. Recent studies show that machine learning models have potential to improve the accuracy of CADx systems. In this work, we present research findings aimed at improving the performance of CADx systems for detecting skin diseases by applying ensemble machine learning models. The investigation encompasses the exploration of three popular classification methods: linear discriminant analysis (LDA), support vector machine (SVM), and convolutional neural network (CNN); and an ensemble model of CNN with SVM. The HAM10000 dataset from Kaggle is used to train and test all classification models. Resampling is employed to address class imbalance in the dataset. Through rigorous experiments, the results highlight the compelling efficacy of the ensemble CNNSVM model, unveiling heightened accuracy up to 92% (from CNN accuracy 85% and SVM accuracy 83%). The outcome of this work has profound implications for artificial intelligence (AI) accelerated medical domains in advancing the accuracy and efficiency of skin disease treatment.
Improving CADx System Performance for Skin Disease Detection using Ensemble Machine Learning Models

Abu Asaduzzaman, Senior Member, IEEE, Christian C. Thompson, and Md J. Uddin

Abstract—Conventional computer-aided diagnosis (CADx) systems play a crucial role in assisting medical professionals with the detection of skin diseases. However, these systems often involve manual, time-consuming, and error-prone processes. Recent studies show that machine learning models have potential to improve the accuracy of CADx systems. In this work, we present research findings aimed at improving the performance of CADx systems for detecting skin diseases by applying ensemble machine learning models. The investigation encompasses the exploration of three popular classification methods: linear discriminant analysis (LDA), support vector machine (SVM), and convolutional neural network (CNN); and an ensemble model of CNN with SVM. The HAM10000 dataset from Kaggle is used to train and test all classification models. Resampling is employed to address class imbalance in the dataset. Through rigorous experiments, the results highlight the compelling efficacy of the ensemble CNN-SVM model, unveiling heightened accuracy up to 92% (from CNN accuracy 85% and SVM accuracy 83%). The outcome of this work has profound implications for artificial intelligence (AI) accelerated medical domains in advancing the accuracy and efficiency of skin disease treatment.

Impact Statement—Although machine learning models such as SVM and CNN are introduced to improve the classification performance of CADx systems, the systems suffer from challenges such as model overfitting and class imbalance that downgrade overall system performance. This work presents an ensemble model of CNN with SVM that inherits the strengths of both CNN and SVM models. To capitalize the benefits of the ensemble CNN-SVM model, we normalize the original HAM10000 dataset for preventing model overfitting and then rectify class imbalances of the dataset. The impact of the proposed ensemble CNN-SVM model includes (i) improved classification performance (up to 92% accuracy) and (ii) enhanced ability of CADx systems to correctly detect and timely treat skin diseases. The proposed machine learning methodology has potential to offer promising AI based solutions for image processing centered medical applications such as treating cancer.

Index Terms—Computer-aided diagnosis (CADx), convolutional neural network (CNN), skin diseases, support vector machine (SVM)

I. INTRODUCTION

Traditionally, the early detection and diagnosis of skin diseases including melanomas have been heavily relied on the expertise of dermatologists through visual examination techniques such as the use of dermatoscopes or invasive procedures such as surgical or needle biopsies [1], [2], [3], [4]. However, these traditional approaches are often time-consuming, expensive, and dependent on the availability of specialized medical professionals. To address these limitations, computer-aided diagnostic (CADx) systems [5], [6], [7] have emerged as valuable means for assisting in the detection of skin malignancies. The objective is to identify dermatological abnormalities in skin images and reduce the number of false positive [8], [9], [10], [11] and false negative [12], [13] readouts by identifying hidden or complex patterns in diagnostic data. A false positive can lead to unnecessary biopsies with many follow-up procedures. On the other hand, a false negative is a wrong diagnosis as it can lead to delays in treatment, making it difficult to treat. In recent years, CADx systems have demonstrated notable advancements in enhancing diagnostic accuracy. However, there are also some potential limitations of CADx systems, such as the possibility of inaccuracy and potential misdiagnosis.

This study aims at improving the performance of CADx systems used for detecting skin diseases using machine learning and deep learning techniques. The use of various machine learning classification algorithms, including LDA, SVM, CNN, and an ensemble model of CNN and SVM is investigated. Generative adversarial network (GAN) is employed to ensure the authenticity of the input dataset. In addition, the dataset is normalized for preventing model overfitting.

This paper is organized as follows. In Section II, background materials are reviewed. Proposed CADx system with machine learning methods, experimental setup, and experimental results are presented in Sections III, IV, and V, respectively. Finally, Section VI concludes the paper.

C. C. Thompson is with Wichita State University, Wichita, KS 67260, USA (e-mail: ccthompson1@shockers.wichita.edu).
M. J. Uddin is with Vanderbilt University, Nashville, TN 37240, USA (e-mail: jashim.uddin@vanderbilt.edu).
II. BACKGROUND MATERIALS

In this section, we review previous work on CADx systems, skin classes, optimal image size, resampling, and GAN.

A. CADx Systems

CADx is a system that uses computer algorithms to assist healthcare professionals in the diagnosis of diseases [14], [15], [16], [17], [18]. They accomplish this by providing clinicians with information about the potential presence of diseases and by analyzing data from medical images like X-rays, magnetic resonance imaging (MRI) and computed tomography (CT) scan. Conventional CADx systems are valuable tools that can improve the accuracy and efficiency of medical diagnosis [19], [20]. They use a sequence of steps to process medical images, including preprocessing, segmentation, feature extraction, feature selection, and classification.

The conventional CADx systems take images for preprocessing. The preprocessing step aims to improve the quality of the medical image by applying various techniques such as noise removal, geometric transformation, cropping, resizing, and adjusting the color balance of the images [21], [22], [23]. The preprocessed image proceeds to the segmentation step. The segmentation step aims to identify and isolate areas of interest within the image [24], [25], [26], [27]. This may entail classifying various tissues and strictures or locating particular areas that are important for a given disease or condition. This is done by identifying regions of similar color, texture, or other features. Feature extraction involves the technique of detecting and extracting the most critical and relevant characteristics from segmented regions [28], [29], [30]. Feature selection involves the process of selecting the most relevant and informative characteristics from a vast set of features derived from medical images [31], [32], [33]. The selected features or regions of interest are used as input in the classification step.

The choice of classification method in a CADx system depends on the trade-offs between accuracy, complexity, and interpretability, as well as the specific requirements of the application at hand. Rule-based algorithms rely on explicit rules and conditions defined by experts. Statistical methods extract quantitative features from images. Machine learning models learn patterns from labeled data to make predictions. In our work, we aim to improve the performance of CADx system by removing fake images, resampling data, and using an ensemble of CNN and SVM models.

B. Skin Classes in the Dataset Used

The Human Against Machine 10000 (HAM10000) dataset is used in this study [34]. HAM10000 dataset, available via Kaggle website, has more than 10,000 training images for detection of pigmented skin lesions with seven classes. The classes are: Melanoma (MEL), Basal Cell Carcinoma (BCC), Vascular Lesion (VAS), Actinic Keratoses (AKIEC), Benign Keratosis-Like Lesions (BKL), Dermatofibroma (DF), and Melanocytic Nevi (NV). MEL develops from pigment-producing cells and is considered the most serious and potentially life-threatening type of skin cancer if not caught early. BCC is the most common type of skin cancer that typically appears as a waxy bump or lesion on the skin. VAS is a variety of skin conditions caused by abnormal blood vessels, including birthmarks and vascular malfunctions. AKIEC is a pre-cancerous lesion that appears as a scaly or crusty growth and is typically caused by sun damage. BKL represents benign skin growths resemble actinic keratoses but have different characteristics. DF is a benign skin lesion that appears as a firm, round bump and is typically brown or reddish-brown. NV is commonly known as moles and is usually benign.

C. Optimal Image Size

The ideal image size for a machine learning classification algorithm depends on several variables, including the complexity of the model, the computing power at hand, the size and complexity of the dataset, and the application’s specific requirements. Larger images generally contain more detailed information, providing the potential for improved feature extraction and better representation of the image characteristics [35], [36]. However, employing larger photos requires more memory and computing power to process and analyze the massive amounts of data in the image. Smaller images demand less memory and processing power, making them more suitable for systems with limited resources. However, reducing the image size may lead to the loss of some crucial details or characteristics of the skin lesions, potentially affecting the classification accuracy [37].

The ideal image size may be discovered empirically by training and testing the model on images of various sizes to see which image size performs the best. In this study, we resized the images from 600 x 450 pixels to 64 x 64 pixels. This adjustment was made to decrease computational demands, expedite training, and mitigate the risk of overfitting.

D. Resampling

Resampling is a statistical approach that uses random extract samples from a dataset to build a new dataset with fewer or more samples that have the same distribution as the original dataset [38], [39], [40], [41]. Resampling is used to generate a more representative sample of the population, improve the performance of a model, or balance an unbalanced dataset. It is particularly useful when dealing with imbalanced datasets, where one class significantly outnumbers the others, leading to potential biases in the model’s predictions.

In this paper, Synthetic Minority Over-sampling Techniques (SMOTE) [42], [43], [44], [45] is used to address the class imbalance issue. By generating these synthetic samples, SMOTE increases the representation of minority classes, thereby reducing the class imbalance and offsetting the negative consequences associated with skewed datasets. The resampled dataset with synthetic samples can then be used to train the machine learning model, enhancing its ability to generalize and make more accurate predictions for both majority and minority classes. By leveraging SMOTE in CADx systems for skin disease detection, one can enhance the model sensitivity to detecting underrepresented skin lesions, leading to more effective and reliable diagnostic outcomes.
E. GAN

GAN [46], [47], [48], [49] is a form of machine learning model that one often uses to generate new data that is similar to the original data. It functions by concurrently training two neural networks: a generator and a discriminator. The generator is in charge of creating fake images that appear like actual images, while the discriminator is in charge of distinguishing between real and fake images. During training, the generator learns to make more realistic images, while the discriminator learns to distinguish between real and fake images, making it more challenging for the generator to produce convincing fakes. GAN is used in this research to eliminate fake images from the dataset. This is because the existence of false images in training dataset can have a major influence on machine learning performance. Fake images can lead to machine learning models learning inaccurate patterns, resulting in poor performance on actual data. By employing GAN, we can distinguish genuine images from potentially misleading ones. This crucial step ensures that our dataset is free from any fake samples, which may otherwise introduce unintended biases and hindered the accuracy of our skin cancer classification model. Nonetheless, it is critical to recognize that training GANs can be computationally costly and necessitates rigorous hyperparameter tuning. Regardless of this limitation, study shows that the use of GAN improves the dependability and effectiveness of skin cancer classification models.

In order to evaluate the significance and effectiveness of GAN in this study, we introduce approximately 20% additional synthetic images for each skin lesion type, amounting to a total of 1,995 fake images in the original HAM10000 dataset (with 10,015 images).

F. Popular Machine Learning Techniques

In this study, three popular machine learning classification techniques (namely, LDA, SVM, and CNN) and an ensemble of CNN and SVM are employed to address skin cancer classification. Each technique brings its own set of advantages and disadvantages as the aim of this experiment is to improve the performance of the CADx system.

1) LDA

LDA is a supervised learning algorithm that is particularly effective for both binary and multi-class classification problems. LDA improves performance by extracting discriminative features and performing effective classification. LDA methods assume that the features follow a normal distribution, which can usually be found in medical images. Furthermore, LDA seeks to enhance the separation between groups, making it particularly effective when there are clear differences between skin lesions. Furthermore, LDA offers a framework for classification, which can be valuable in clinical decision making. However, LDA also comes with a few limitations. One such limitation is that it assumes that classes have the same covariance matrices, which may not always be true in complex medical imaging situations. In CADx systems, LDA can be utilized as a classifier to distinguish between different classes of medical conditions or abnormalities based on the extracted features.

2) SVM

SVM methods are very effective in handling high-dimensional features, which are common in medical image datasets. SVMs can efficiently separate classes even when the data is not linearly separable through the use of different kernel functions, allowing them to capture complex relationships. Additionally, SVMs inherently incorporate the concept of margin, aiming to maximize the separation between classes and thus potentially leading to better generalization performance. However, SVMs also come with a few drawbacks. They can be computationally intensive, especially with large datasets, which may impact their real-time applicability in clinical settings. The choice of the kernel and its associated parameters can significantly influence the model’s performance, and finding the optimal combination may require extensive tuning. SVMs can also be sensitive to the choice of hyperparameters, and improper configuration may lead to suboptimal results. SVM may improve CADx system performance by providing effective classification, robustness to overfitting, some degree of interpretability, and scalability.

3) CNN

CNN models excel at extracting the right features from images, eliminating the need for manual feature engineering. This capability is crucial in medical imaging, where intricate patterns and subtle details are vital for accurate diagnosis. CNNs are also capable of handling large, high-resolution images well, making them suitable for skin analysis. Additionally, CNNs are adept at capturing spatial relationships in data, allowing them to discern patterns that may be challenging for traditional machine learning models. However, CNNs come with a few limitations. They demand a lot of computing resources, especially when dealing with deep architectures or big data such as the HAM10000 dataset. Training a CNN from scratch is time-consuming, and fine-tuning pre-trained models may still require significant computational power. CNN models have potential to improve CADx system performance by automatically learning hierarchical features, exploiting spatial hierarchies, and being robust to variability.

4) Ensemble Model of CNN and SVM

Combining CNN and SVM models offers a powerful approach for skin cancer classification. This combination leverages the strengths of both models. By combining the strengths of CNN and SVM, the ensemble model can achieve higher classification accuracy compared to the individual models. Moreover, this approach provides redundancy and robustness against overfitting, as the two models are inherently different in their learning processes and feature extraction methods. However, implementing the ensemble model requires careful tuning and integration, and the computational resources needed for training are substantial. Ensuring compatibility and coherence between the CNN and SVM components is crucial, and finding the right balance between their contributions can be a complex process. Nevertheless, with proper optimization and validation, an ensemble of CNN and SVM can offer a powerful tool for accurate skin cancer classification.
III. MACHINE LEARNING MODELS FOR CADX SYSTEMS

In this section, we present a comprehensive description of the machine learning-based classification techniques proposed for CADx systems to improve performance. Figure 1 illustrates the major steps of an improved CADx system, where Detect & Remove Fake Images, Normalize Dataset, and Resample to Avoid Imbalance are the newly introduced steps to typical CADx systems.

Figure 1. Workflow diagram of the proposed methodology to reduce the number of input data fields

A. Detect and Remove Fake Images

To address dataset authenticity concerns and enhance the reliability of the CADx system, GAN technique is employed in the proposed system as a promising tool for validating the dataset. GAN is leveraged to scrutinize the dataset for any fake generated images. By subjecting the dataset to GAN-based analysis, the CADx system can effectively identify and filter out any inauthentic samples, ensuring that the model is not trained on misleading or erroneous data. This validation process helps maintain the integrity of the dataset and improves the CADx system’s ability to accurately diagnose and classify skin lesions, contributing to more trustworthy results in dermatological research.

B. Normalize the Input Dataset

To analyze and normalize the original dataset for preventing machine learning overfitting, we introduce exploratory data analysis (EDA) in the proposed CADx system. We start with 10,015 images from the HAM10000 dataset. We plan to assimilate approximately 20% newly generated (fake) images. This brings the total number of images to 12,010. Because the HAM10000 dataset has seven types, we generate 285 new images for each type using GAN. Table I shows the result of EDA on the dataset. This finding shows the differences among various types. Utilizing these images enable machine learning models to be tailored and trained to account the distinctive traits and patterns displayed by each group.

C. Preprocessing

After addressing fake images and normalizing dataset, preprocessing serves as the next stage of the proposed CADx system with machine learning, wherein skin lesion images from HAM10000 dataset undergo enhancement and preparation for further analysis. Techniques like image resizing and noise removal are implemented to ensure that standardized image formats are compatible with machine learning algorithms.

D. Resample to Avoid Imbalance

Addressing class imbalances in the dataset is crucial for accurate skin lesion classification. To rectify this, resampling techniques like SMOTE are employed. SMOTE creates synthetic samples for the minority class by interpolating...
existing ones, balancing the dataset. Table II illustrates the distribution of images within the dataset after using SMOTE technique. By doing so, the CADx system can learn from a more diverse and representative dataset, reducing overfitting and improving generalization.

<table>
<thead>
<tr>
<th>Diagnostic Category</th>
<th>ORIGINAL + 285 NEW/FAKE IMAGES</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Melanoma (MEL)</td>
<td>900</td>
<td>14.28</td>
</tr>
<tr>
<td>Basal Cell Carcinoma (BCC)</td>
<td>900</td>
<td>14.28</td>
</tr>
<tr>
<td>Vascular Lesion (VAS)</td>
<td>900</td>
<td>14.28</td>
</tr>
<tr>
<td>Actinic Keratoses (AKIEC)</td>
<td>900</td>
<td>14.28</td>
</tr>
<tr>
<td>Benign Keratosis-Like Lesions (BKL)</td>
<td>900</td>
<td>14.28</td>
</tr>
<tr>
<td>Dermatofibroma (DF)</td>
<td>900</td>
<td>14.28</td>
</tr>
<tr>
<td>Melanocytic Nevi (NV)</td>
<td>900</td>
<td>14.28</td>
</tr>
<tr>
<td>Total</td>
<td>6300</td>
<td>~100.00</td>
</tr>
</tbody>
</table>

### E. Segmentation

Segmentation in the proposed CADx system involves the isolation of the skin lesion region from the background and other non-lesion areas within the image. By accurately segmenting the relevant areas for classification (i.e., thresholding the image), the CADx system enhances precision while reducing computational complexity. Figure 3 shows an example of thresholding used in segmentation.

![Segmentation example](image)

### F. Feature Extraction and Selection

Subsequent to segmentation, feature extraction and selection is done for the CADx system. Relevant features are extracted from the skin lesion region, encompassing texture, color, shape, and other discriminative characteristics unique to different skin lesions. Feature selection methods are then applied to identify the most informative features, optimizing model performance by reducing dimensionality.

### G. Classification

The classification stage in the CADx system is a pivotal step that heavily influences its overall performance and accuracy in skin disease detection. By employing popular machine learning algorithms such as LDA, SVM, CNN, and an ensemble CNN-SVM model, it is expected that the CADx system achieves improved classification accuracy. These algorithms are designed to handle complex patterns and relationships within the data, enabling the model to discern intricate differences between various skin lesion types. SVM are effective in handling high-dimensional data, while CNN excels at capturing spatial features and hierarchical patterns in images. Ensemble methods combine the strengths of CNN and SVM, leveraging the power of both techniques to achieve better generalization and robustness. By leveraging these advanced classification techniques, the CADx system enhances its ability to accurately diagnose and classify diverse skin lesions, providing valuable support to dermatologists and contributing to more effective medical interventions.

### IV. EXPERIMENTAL SETUP

In this section, we describe experimental details including HAM10000 dataset used to train and test the classification models, hyperparameters of the classification algorithms, and the shape of the CNN algorithm.

#### A. Dataset Used: HAM10000

The HAM10000 dataset consists of a large number of dermatoscopic images of common pigmented skin lesions [34]. It is created by Philipp Tschantl, Cliff Rosendahl, and Harald Kittler at the Medical University of Graz in Austria [50], [51]. The dataset consists of more than 10,000 dermatoscopic images (in seven types) of pigmented skin lesions, which have been manually labeled by expert dermatologists. Table III shows the seven diagnostic categories of skin diseases within the HAM10000 dataset. In this study, we use 10,015 original HAM10000 images and assimilate approximately 20% additional synthetic images.

#### B. Classification Model Architecture

We explore four classification models (i.e., LDA, SVM, CNN, and ensemble CNN-SVM) in the proposed CADx system to enhance skin lesion classification performance. For the classification algorithms, we strategically consider vital hyperparameters to optimize each model’s performance. Table IV summarizes the hyperparameters of the classification algorithms. For LDA, the selection of solver, the application of shrinkage, and the determination of the number of components are pivotal in shaping classification capabilities. In this work, we use singular value decomposition (SVD) solver. Likewise, for SVM, careful parameterization focused on kernel selection, the fine-tuning of the regularization parameter (C), and the degree of polynomial for optimal decision boundaries are pivotal. In this work, we use kernel = poly, C = 1.0, and degree of polynomial = 3. The CNN design encompasses key hyperparameters such as the number of convolutional layers (4), dropout layers (4), and hidden layers (2), filter size (4, 3), max pooling pool size (2, 2), dropout rate (25%), activation function (ReLU), L2 regularization rate (0.1%), and training epochs (65). In crafting our ensemble CNN-SVM model, we emphasize on the weighting mechanism for inconsistent predictions (70% for CNN and 30% for SVM) to achieve a harmonious fusion of these two powerful algorithms.

Next we discuss the CNN model architecture in detail as it is specifically designed for image classification. Table V shows the shape and number of parameters of the CNN classification algorithm. The CNN model consists of four convolutional layers with filter sizes of (4, 3) in each layer, facilitating the extraction of hierarchical features from the input images.
TABLE III  
Diagnostic Categories of HAM10000 Dataset

<table>
<thead>
<tr>
<th>Category</th>
<th>ORIGINAL IMAGE</th>
<th>Fake Image</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEL</td>
<td>![Image 1]</td>
<td>![Image 2]</td>
</tr>
<tr>
<td>BCC</td>
<td>![Image 3]</td>
<td>![Image 4]</td>
</tr>
<tr>
<td>VAS</td>
<td>![Image 5]</td>
<td>![Image 6]</td>
</tr>
<tr>
<td>AKIEC</td>
<td>![Image 7]</td>
<td>![Image 8]</td>
</tr>
<tr>
<td>BKL</td>
<td>![Image 9]</td>
<td>![Image 10]</td>
</tr>
<tr>
<td>DF</td>
<td>![Image 11]</td>
<td>![Image 12]</td>
</tr>
<tr>
<td>NV</td>
<td>![Image 13]</td>
<td>![Image 14]</td>
</tr>
</tbody>
</table>

After each convolutional layer, max pooling layers with a pool size of (2, 2) are employed to down sample the extracted features, aiding in reducing spatial dimensions. To mitigate overfitting, dropout layers with a dropout rate of 0.25 are strategically inserted after each max pooling layer. The depth of the feature maps increases progressively through the network, starting with 32 filters in the first layer and reaching 256 filters in the final convolutional layer. A global average pooling layer is incorporated to transform the spatial dimensions into a vector of length 256, contributing to reducing the total number of parameters. After the convolutional layers, there are two fully connected dense layers with 127 and 7 units, respectively. The first dense layer employs Rectified Linear Unit (ReLU) activation and L2 regularization with a rate of 0.001, enhancing the model’s ability to learn intricate patterns in data. The entire model comprises a total of 550,759 parameters, which include weights and biases. During training, the CNN model undergoes 65 epochs with early stopping and learning rate reduction callbacks, ensuring effective convergence and preventing overfitting of the CNN model on the training data.

V. RESULTS AND DISCUSSION

This section presents experimental results obtained by simulating the proposed CADx system to assess the impact of
the ensemble model of CNN and SVM on classifying skin disease images. First, we discuss the performance of a typical CADx system without removing fake images and without applying resampling technique. Then we assess the impact of removing fake images and resampling the dataset on the performance of LDA, SVM, CNN, and the ensemble CNN-SVM models.

A. Typical CADx System Performance

The first set of experimental results is obtained by utilizing images from the original HAM10000 dataset plus an additional 20% of generated fake images. So, the input dataset is the cumulation of 10,015 samples from the HAM10000 dataset and 1,995 (285x7) fake images.

The classification performance (as F1 score and accuracy) of LDA, SVM, CNN, and the ensemble CNN and SVM without GAN is presented in Table VI. The models exhibit notable disparities in their performance across different skin lesion classes. Particularly noteworthy is the poor performance for the DF class, where F1 score is reported as 0.00 for LDA and CNN. This indicates significant challenges in correctly identifying and classifying instances of DF, highlighting an area where the model may need improvement. In contrast, the models perform relatively well for the NV class, with a high F1 score for all classification models. It is important to notice that the ensemble CNN-SVM model exhibits prominent strengths in accurately classifying the majority of skin lesions, particularly achieving high F1 score for the NV class with values of 0.95. This contributes to an impressive overall accuracy of 0.79. However, there are areas (such as detecting and excluding fake images from the input dataset and applying resampling technique) for improvement, especially in handling less prevalent classes like DF and AKIEC.

<table>
<thead>
<tr>
<th>Category</th>
<th>LDA</th>
<th>SVM</th>
<th>CNN</th>
<th>Ensemble Model (CNN-SVM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEL</td>
<td>0.21</td>
<td>0.29</td>
<td>0.40</td>
<td>0.40</td>
</tr>
<tr>
<td>BCC</td>
<td>0.20</td>
<td>0.42</td>
<td>0.54</td>
<td>0.52</td>
</tr>
<tr>
<td>VAS</td>
<td>0.16</td>
<td>0.48</td>
<td>0.68</td>
<td>0.69</td>
</tr>
<tr>
<td>AKIEC</td>
<td>0.16</td>
<td>0.28</td>
<td>0.33</td>
<td>0.20</td>
</tr>
<tr>
<td>BKL</td>
<td>0.19</td>
<td>0.36</td>
<td>0.49</td>
<td>0.44</td>
</tr>
<tr>
<td>DF</td>
<td>0.00</td>
<td>0.19</td>
<td>0.00</td>
<td>0.22</td>
</tr>
<tr>
<td>NV</td>
<td>0.72</td>
<td>0.83</td>
<td>0.88</td>
<td>0.95</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.52</td>
<td>0.67</td>
<td>0.76</td>
<td>0.79</td>
</tr>
</tbody>
</table>

B. CADx System Performance without Fake Data

The second set of experimental results is obtained by utilizing 12,010 images (10,015 original HAM10000 images plus 1,995 fake images) with the use of GAN to detect and exclude the fake images. The F1 scores and accuracy of LDA, SVM, CNN, and the ensemble SVM-CNN models with GAN are presented in Table VII. It is notable that the performance of the DF class has improved significantly, where F1 score is reported as the highest (0.70) for the LDA model. Across various skin lesion classes, the LDA model demonstrates relatively balanced performance, the SVM model showcases robust performance, and the CNN model exhibits outstanding performance. As expected, the ensemble CNN-SVM model demonstrates the best performance across various skin lesion classes indicating its effectiveness in classification.

<table>
<thead>
<tr>
<th>Category</th>
<th>LDA</th>
<th>SVM</th>
<th>CNN</th>
<th>Ensemble Model (CNN-SVM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEL</td>
<td>0.57</td>
<td>0.70</td>
<td>0.71</td>
<td>0.79</td>
</tr>
<tr>
<td>BCC</td>
<td>0.74</td>
<td>0.82</td>
<td>0.87</td>
<td>0.92</td>
</tr>
<tr>
<td>VAS</td>
<td>0.92</td>
<td>0.97</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>AKIEC</td>
<td>0.82</td>
<td>0.88</td>
<td>0.87</td>
<td>0.95</td>
</tr>
<tr>
<td>BKL</td>
<td>0.60</td>
<td>0.69</td>
<td>0.70</td>
<td>0.80</td>
</tr>
<tr>
<td>DF</td>
<td>0.89</td>
<td>0.96</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>NV</td>
<td>0.64</td>
<td>0.73</td>
<td>0.75</td>
<td>0.83</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.75</td>
<td>0.83</td>
<td>0.85</td>
<td>0.92</td>
</tr>
</tbody>
</table>

C. CADx System Performance without Fake Data and with Resampled Dataset

The third set of experimental results is obtained for the CADx system by resampling the dataset without fake images to 900 samples per category. The F1 scores and accuracy of LDA, SVM, CNN, and the ensemble CNN-SVM models are presented in Table VIII. Across various skin lesion classes, all models show significant improvement in F1 score and accuracy. It is remarkable that the ensemble CNN-SVM model shows the best performance (with a F1-score of 0.99 for the VAS samples and an accuracy of 0.92).

<table>
<thead>
<tr>
<th>Category</th>
<th>LDA</th>
<th>SVM</th>
<th>CNN</th>
<th>Ensemble Model (CNN-SVM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEL</td>
<td>0.83</td>
<td>0.98</td>
<td>0.91</td>
<td>0.92</td>
</tr>
<tr>
<td>BCC</td>
<td>0.72</td>
<td>0.87</td>
<td>0.89</td>
<td>0.91</td>
</tr>
<tr>
<td>VAS</td>
<td>0.92</td>
<td>0.97</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>AKIEC</td>
<td>0.82</td>
<td>0.88</td>
<td>0.87</td>
<td>0.95</td>
</tr>
<tr>
<td>BKL</td>
<td>0.60</td>
<td>0.69</td>
<td>0.70</td>
<td>0.80</td>
</tr>
<tr>
<td>DF</td>
<td>0.89</td>
<td>0.96</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>NV</td>
<td>0.64</td>
<td>0.73</td>
<td>0.75</td>
<td>0.83</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.75</td>
<td>0.83</td>
<td>0.85</td>
<td>0.92</td>
</tr>
</tbody>
</table>

The ensemble CNN-SVM model in the CADx system with the incorporation of GAN and the utilization of resampled dataset, has demonstrated substantial improvement (with respect to F1-score and accuracy) in skin lesion classification across various contemporary models. The improvement emphasizes the effectiveness of the ensemble CNN-SVM model in CADx systems for detecting skin diseases.

VI. CONCLUSIONS

This study addresses the challenges inherent in skin lesion classification by introducing promising machine learning techniques such as ensemble models. We implement three popular image classification models: LDA, SVM, and CNN, and introduce an ensemble CNN-SVM model. To exploit the performance of the ensemble CNN-SVM model, we use GAN (for removing fake images from the input dataset) and
ensemble models in our future endeavor. Initially, for a typical CADx system without removing fake images and without applying resampling technique, ensemble CNN-SVM model accuracy is (0.79), followed by CNN (0.76), SVM (0.67), and LDA (0.52). The incorporation of GAN effectively addresses challenges associated with fake images, leading to improvements in overall accuracy and F1 score across all classification algorithms. The resampled dataset consistently outperformed the input dataset with and without GAN, emphasizing the critical role of addressing class imbalance in facilitating accurate predictions for all skin lesion classes. Finally, the ensemble CNN-SVM model exhibits the highest accuracy of 0.92, followed by CNN (0.84), SVM (0.83), and LAD (0.75%).

We plan to explore the integration of advanced machine learning techniques, the exploration of additional data processing methods, and the development of more sophisticated ensemble models in our future endeavor.

REFERENCES


Dominion University in Virginia, USA, the International Society for Engineering Research and Development in Thailand, and the IEEE Wichita Professional Section in Kansas, USA. Dr. Asaduzzaman is a life-time member of several honor societies including Phi Kappa Phi (PKP), Tau Beta Pi (TBP), Upsilon Pi Epsilon (UPE), and Golden Key. He received IEEE Region 5 Outstanding Student Branch Counselor Award in 2018. He has been serving as IEEE Region 5 Executive Member and IEEE Wichita Section Officer since 2019.

Christian C. Thompson earned his master’s degree in data science from Wichita State University (WSU), Kansas, USA in December 2023. He obtained his bachelor’s degree in aerospace engineering with minors in computer science and mathematics from WSU in May 2022. As Graduate Research Assistant at WSU, Mr. Thompson has applied the optical character recognition (OCR) techniques and advanced sequence modeling to efficiently extract data, streamlining workflows. He has developed custom Python scripts for automated data wrangling, contributing to a significant reduction in preprocessing time. He has used Tableau to create comprehensive dashboards for visualizing sales, customer behavior, and business performance metrics. As Data Management Intern at Cotiviti: Healthcare Analytics Company, he has leveraged the Federal Information Processing Standards (FIPS) codes from the U.S. Census Bureau for precise geographic identification. As Student Assistant at the National Institute for Aviation Research (NIAR) at WSU, he engineered a sophisticated C++ codebase to detect and track trajectories of small, unmanned aircraft systems, resulting in a notable reduction in collision risks. Mr. Thompson is a member of the National Society of Black Engineers (NSBE), Tau Beta Pi (TBP), the American Institute of Aeronautics and Astronautics (AIAA), and the American Society for Engineering Education (ASEE).

Md. J. Uddin received his PhD degree in organic chemistry from the Shinshu University School of Engineering, Ueda, Japan in 2001. He obtained his MS degree in organic chemistry and BS Honors degree in chemistry from the University of Dhaka, Bangladesh in 1993 and 1991, respectively. Currently, Dr. Uddin is an Associate Professor of Research and primary faculty member (academic) in the Department of Biochemistry at Vanderbilt University, Tennessee, USA. He has authored 55 peer-reviewed scientific publications in the leading molecular imaging, organic chemistry, bioorganic chemistry, and medicinal chemistry journals, four issued United States Patents, and two book chapters. He authored two cover articles, one in Cancer Pretension Research (2011) and the other in ACS Medicinal Chemistry Letters (2014). He is the Principal Investigator (PI) of two nationally prestigious research grants from the Phi Beta Psi Trust (AWD00000652 and AWD0001248, 8/2022 –7/2024). In addition, he is the Multiple Principal Investigator (MPI) of a federal research grant that he has received from the National Institutes of Health, National Cancer Institute (NIH/NCI R01 CA260958-01A1, 7/2021-6/2026) for the development of a gene therapy for triple-negative breast cancer. Dr. Uddin served as the Chair of the Network of Translational Research of the National Cancer Institute during 7/2011-6/2012. He serves as member of several professional societies including the American Chemical Society, the Society for Pharmaceutical Chemists Bangladesh, the Japanese Chemical Society, and the Canadian Society for Pharmaceutical Sciences.