Skin Cancer Classification Using Machine Learning and Neural Networks

CSAI 403 Machine Learning and Pattern Recognition Project Report

Team Members: [Insert Team Member Names]
Course: CSAI 403 - Machine Learning and Pattern Recognition
Institution: The British University in Dubai
Date: June 2025

Abstract

Skin cancer represents one of the most dangerous forms of cancer world wide with early detection being important for successful treatment results. This project presents a complete machine learning approach to automated skin cancer classification using the HAM10000 dataset. We developed and evaluated multiple machine learning models which including traditional algorithms and neural networks to classify seven different types of skin lesions. Our study shows that convolutional neural networks (CNN) achieve best performance with 95.03% accuracy which performed better than traditional machine learning approaches. Additionally, we implemented a graphical user interface (gui) using streamlit to provide practical accessibility to our trained model.

Contents

1. Introduction

The increasing cases of skin cancer worldwide needs the development of an diagnostic tools that can help healthcare professionals in early detection and classification. Dermoscopic image analysis has appeared as a important component in dermatological diagnosis with manual interpretation requires more expertise and is subject to inter-observer variability. Machine learning techniques which includes particularly deep learning approaches have shown potential in medical image analysis tasks.

Our project shows this challenge by developing a complete machine learning system which capable of classifying skin lesions into seven distinct categories: Melanocytic nevi, Melanoma, Benign keratosis-like lesions, Basal cell carcinoma, Actinic keratoses, Vascular lesions and Dermatofibroma. We approached this as both a discriminative modeling task which focusing on accurate classification and explored various algorithmic approaches to understand their strengths and limitations.

The importance of this work extends beyond academic exploration as accurate automated classification systems could serve as important screening tools in clinical settings which particularly in resource limited environments where specialist dermatologists may not be readily available. Our implementation includes a practical user interface that shows the real world applicability of our trained models.

2. Dataset Description and Exploratory Data Analysis

We uses the ham10000 (Human Against Machine with 10,000 training images) dataset, which represents one of the largest publicly available collections of dermatoscopic images. This dataset contains 10,015 dermatoscopic images across seven diagnostic categories which making it an ideal choice for multi-class classification tasks in dermatological applications.

2.1 Dataset Characteristics

The dataset shows significant class imbalance which is characteristic of real world medical datasets. Our exploratory data analysis shows that melanocytic nevi comprises the largest class with 6,705 samples while Dermatofibroma represents the smallest class with only 142 samples. This imbalance established challenges for model training and required careful consideration in our preprocessing pipeline.

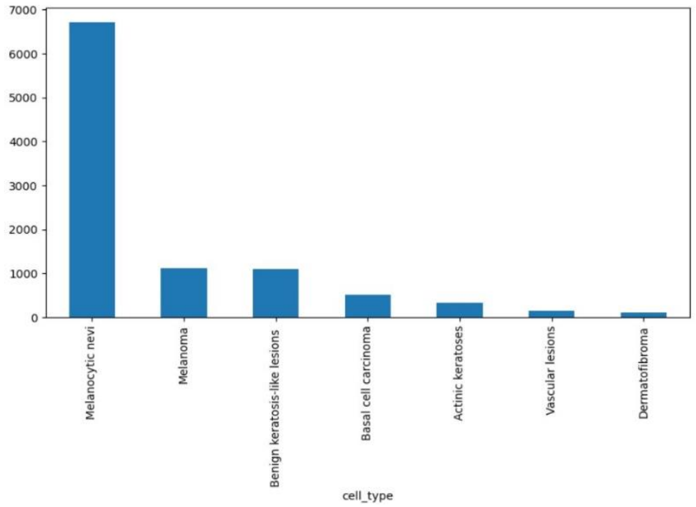
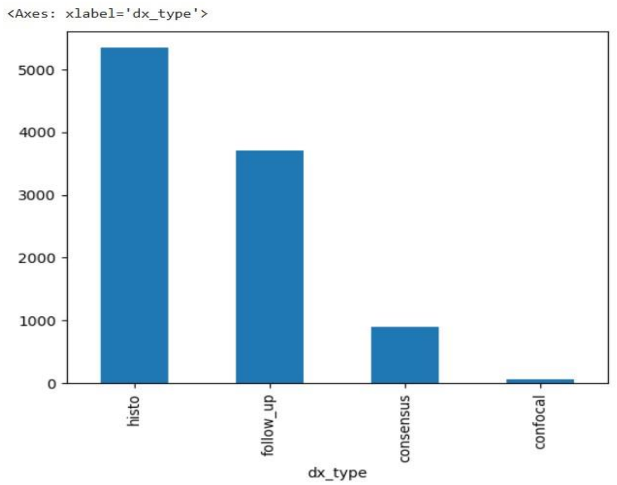
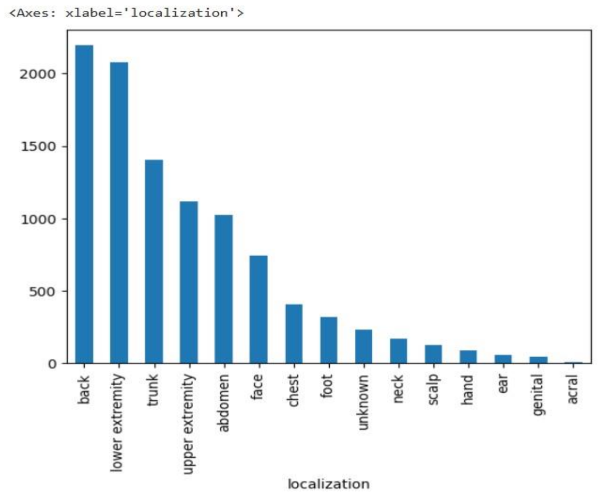


FIGURE 1: Class Distribution Bar Chart - showing cell_type counts

The demographic distribution shows a predominance of male patients (54.4%) over female patients (45.4%) with a small percentage of unknown gender cases. Patient ages follow a normal distribution with a mean of approximately 52 years which ranging from infants to elderly patients. Lesion localization analysis shows that the back and lower extremities are the most common sites which followed by the trunk and upper extremities.



```
<Axes: xlabel='dx_type'>
```



```
<Axes: xlabel='localization'>
```

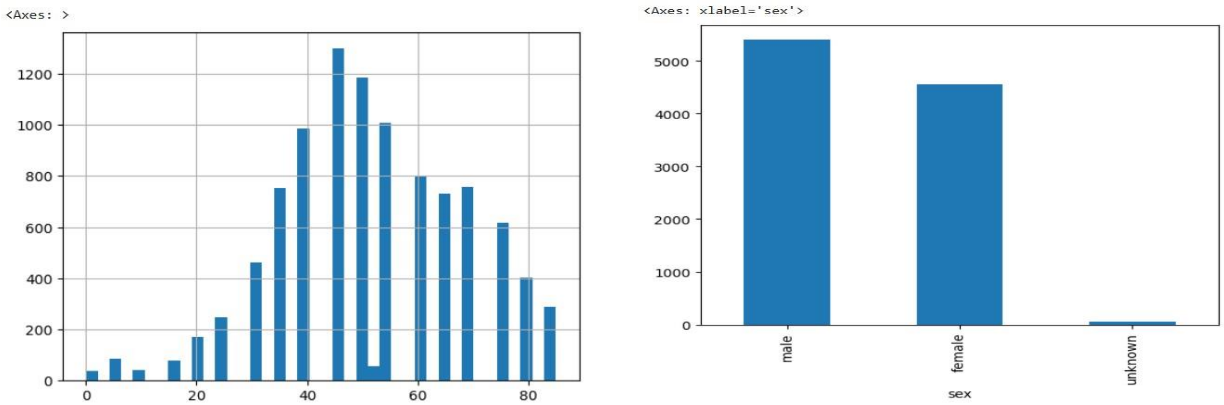


FIGURE 2: Demographic Analysis Charts - showing dx_type, localization, age histogram, and sex distribution

2.2 Data Preprocessing and Cleaning

We implemented a complete preprocessing pipeline to confirm data quality and model compatibility. Missing age values were imputed using the dataset mean which maintaining consistency while saving the overall age distribution. All images were standardized to 128×128 pixels to confirm the computational efficiency while retaining sufficient detail for classification tasks.

Our preprocessing function normalized pixel values to the range (0,1) which is important for neural network training stability. This normalization approach confirms that all input features contribute equally to the learning process and prevents gradient explosion during backpropagation.

3. Data Balancing Strategy

The serious class imbalance in our dataset required careful attention to prevent model bias toward majority classes. We applied smote (Synthetic Minority Oversampling Technique) to solve this challenge. Smote generates synthetic samples for minority classes by introducing between existing samples in the feature space which effectively creating a balanced dataset without simply duplicating existing samples.

After applying smote, each class contained exactly 6705 samples which resulting in a perfectly balanced dataset of 46,935 total samples. This approach improved model performance across all classes. The balanced dataset confirms that our models learn to recognize patterns specific to each skin lesion type rather than simply predicting the majority class.

4. Methodology and Model Implementation

4.1 Baseline Machine Learning Models

We implemented four machine learning models to prove baseline performance metrics. Each algorithm was chosen to represent different learning paradigms and to provide information into the nature of the classification problem.

Decision Tree Classifier: We configured our decision tree with a maximum depth of 10 and minimum samples split of 10 to prevent overfitting while maintaining model interpretability. Decision trees are particularly important in medical applications due to their explainable nature which allowing healthcare professionals to understand the decision making process.

Naive Bayes Classifier: The gaussian naive bayes implementation believes feature independence which often violated in image data that provides a probabilistic baseline for comparison. This algorithm simplicity makes it efficient and helps as an important benchmark.

Logistic Regression: We applied logistic regression with LBFGS optimization and increased iteration limit to confirm convergence. This linear model provides information into the linear separability of our feature space when images are flattened to vectors.

K-Nearest Neighbors (K-NN): With k=5 this learning algorithm classifies samples based on the majority class among their nearest neighbors. Knn performance on our flattened image data provides information into the local structure of the feature space.

4.2 Neural Network Architectures

Multi-Layer Perceptron (MLP): Our MLP architecture consists of four fully connected layers with 512, 256 and 128 neurons respectively which includes in a 7 neuron output layer with softmax activation. We applied batch normalization after each hidden layer to stabilize training and reduce internal covariate shift. Dropout layers with rates of 0.5, 0.5, and 0.3 provide regularization to prevent overfitting.

The MLP flattens input images into 49,152 dimensional vectors by learning complex non-linear mappings between pixel intensities and class labels. While conceptually simpler than convolutional architectures MLP can capture global patterns across the entire image.

Convolutional Neural Network (CNN): Our cnn architecture applies the spatial structure essential in image data. The network consists of four convolutional blocks, each containing a convolutional layer, batch normalization and max pooling. Filter sizes progress from 32 to 256 which allowing the network to learn increasingly complex features from low level edges to high-level semantic patterns.

The convolutional layers use 3×3 kernels with relu activation which is a standard configuration that has effective in image classification tasks. Max pooling layer with 2×2 windows reduces spatial dimensions while retaining the most relevant features. The final fully connected layers combine these learned features for classification decisions.

5. Training Methodology

We split our balanced dataset using an 80 20 train test division by stratified to maintain class balance across both sets. The training set contained 37548 samples while the test set contained 9387 samples. For neural networks, we further allocated 20% of the training data for validation which enabling us to monitor training progress and implement early stopping if necessary.

Our training configuration applied the adam optimizer with an initial learning rate of 0.001, chosen for its adaptive learning rate properties and robust performance across various optimization. We implemented reduce lr on plateau callbacks to automatically reduce the learning rate when validation loss plateaued which preventing training immobility.

Batch sizes of 32 provided an effective balance between gradient estimate quality and computational efficiency. We trained neural networks for a maximum of 20 epochs by the learning rate reduction mechanism allowed for effective training within this timeframe.

6. Results and Performance Analysis

6.1 Baseline Model Performance

Our baseline models shows varying levels of effectiveness which shows important information about the nature of skin lesion classification. The performance metrics across all baseline models are as below:

- **Decision Tree:** achieved 73.68% accuracy with an f1-score of 73.45% which providing decision paths while maintaining reasonable classification performance
- **Naive Bayes:** performed very poorly with only 38% accuracy which highlighting the violation of feature independence in image data
- **Logistic Regression:** significantly performed better than other linear methods with 92.43% accuracy and 92.25% f1-score
- **K-Nearest Neighbors:** delivered the best baseline performance with 93.84% accuracy and 93.28% f1-score

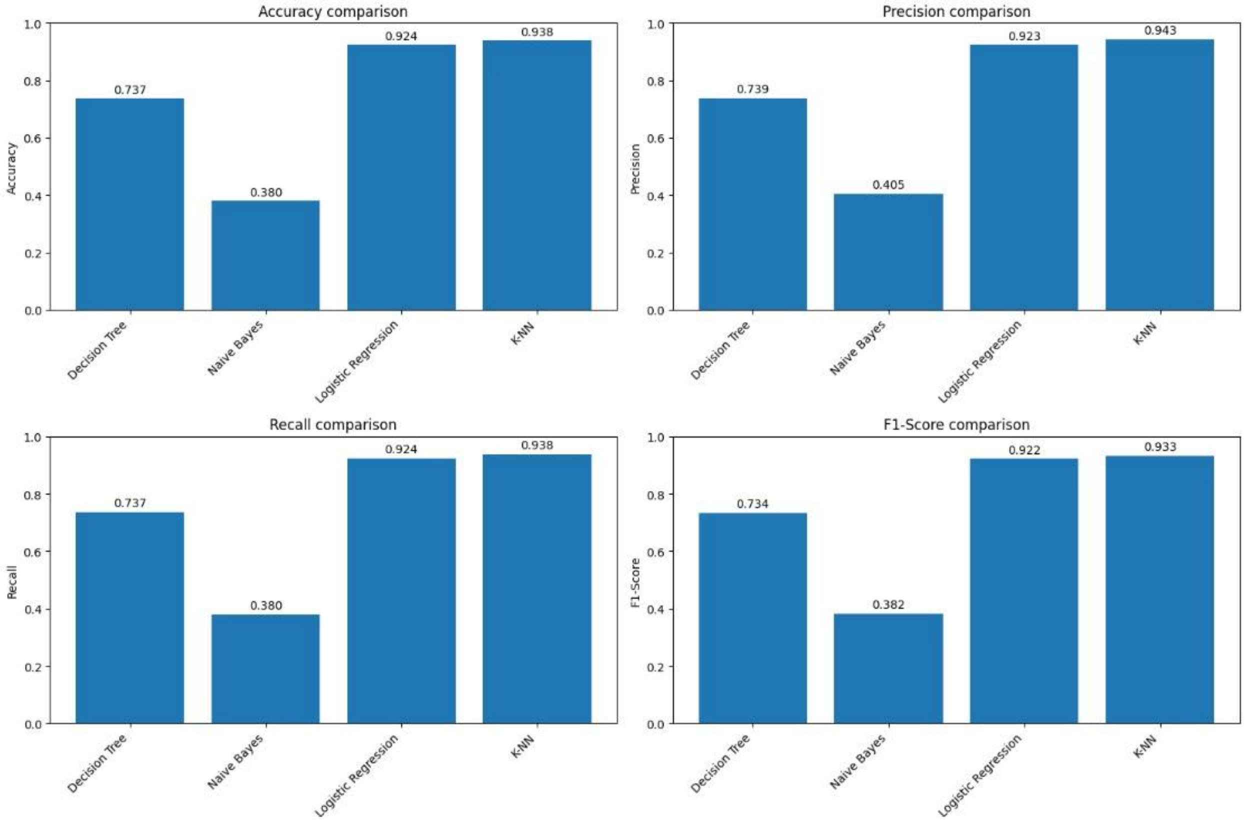
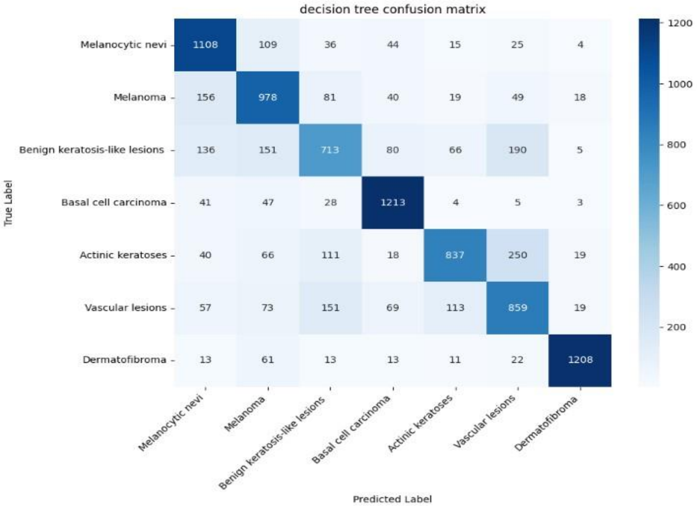


FIGURE 3: Baseline Models Comparison Chart - showing accuracy, precision, recall, and F1-score for all four models

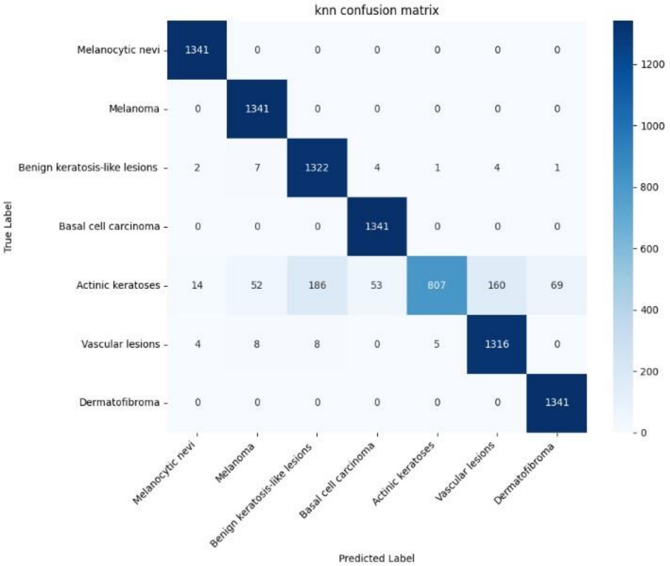
The Decision Tree moderate performance shows that pixel based features contain important discriminative information for reasonable classification accuracy while maintaining model interpretability. Naive Bayes face difficulties due to the large correlation between neighboring pixels which making the independence assumption particularly problematic for this algorithm.



	Melanocytic nevi	Melanoma	Benign keratosis-like lesions	Basal cell carcinoma	Actinic keratoses	Vascular lesions	Dermatofibroma
Melanocytic nevi	1108	109	36	44	15	25	4
Melanoma	156	978	81	40	19	49	18
Benign keratosis-like lesions	136	151	713	80	66	190	5
Basal cell carcinoma	41	47	28	1213	4	5	3
Actinic keratoses	40	66	111	18	837	250	19
Vascular lesions	57	73	151	69	113	859	19
Dermatofibroma	13	61	13	13	11	22	1208

FIGURE 4: Decision Tree Confusion Matrix

Logistic Regression strong performance shows that linear combinations of pixel features can effectively separate many skin lesion classes though some non linear patterns remain uncaptured. K-nearest Neighbors better performance shows that the balanced dataset creates well defined clusters in the high dimensional pixel space which allowing similarity based classification to succeed effectively.



	Melanocytic nevi	Melanoma	Benign keratosis-like lesions	Basal cell carcinoma	Actinic keratoses	Vascular lesions	Dermatofibroma
Melanocytic nevi	1341	0	0	0	0	0	0
Melanoma	0	1341	0	0	0	0	0
Benign keratosis-like lesions	2	7	1322	4	1	4	1
Basal cell carcinoma	0	0	0	1341	0	0	0
Actinic keratoses	14	52	186	53	807	160	69
Vascular lesions	4	8	8	0	5	1316	0
Dermatofibroma	0	0	0	0	0	0	1341

FIGURE 5: KNN Confusion Matrix

6.2 Neural Network Performance

The multi layer perceptron with its architecture achieved only 14.32% accuracy with an f1-score of 3.64%. This disappointing performance likely results from the curse of dimensionality when flattening 128×128×3 images into 49,152-dimensional vectors. The training curves showed early overfitting which suggesting that the MLP face difficulties to generalize from the high-dimensional input space.

In contrast our convolutional neural network achieved good performance with 95.03% accuracy and 94.91% f1-score. The cnn ability to apply spatial relationships and hierarchical feature learning proved important for skin lesion classification. The training history shows steady convergence with validation accuracy reaching over 96% during training.

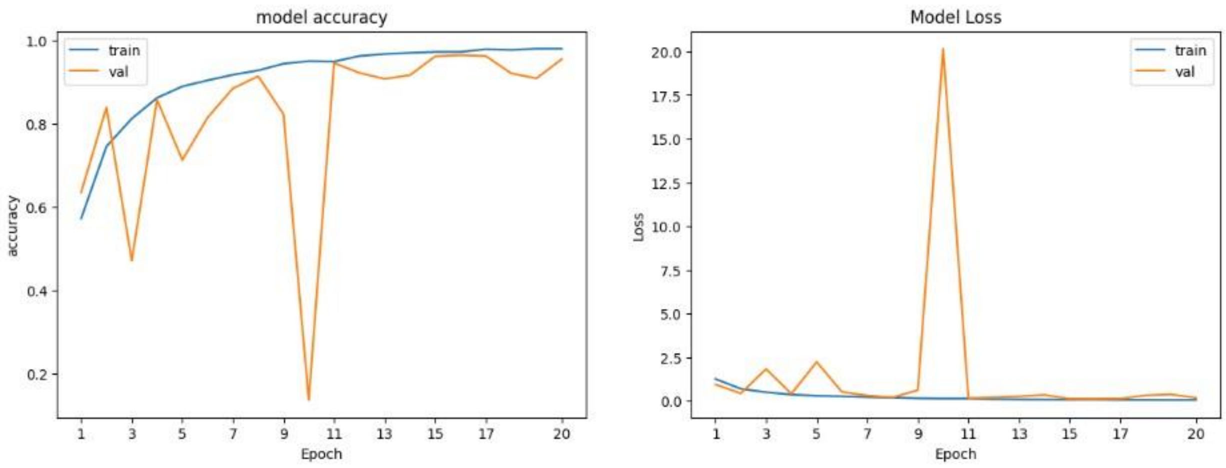


FIGURE 6: CNN Training History - showing model accuracy and loss curves over epochs

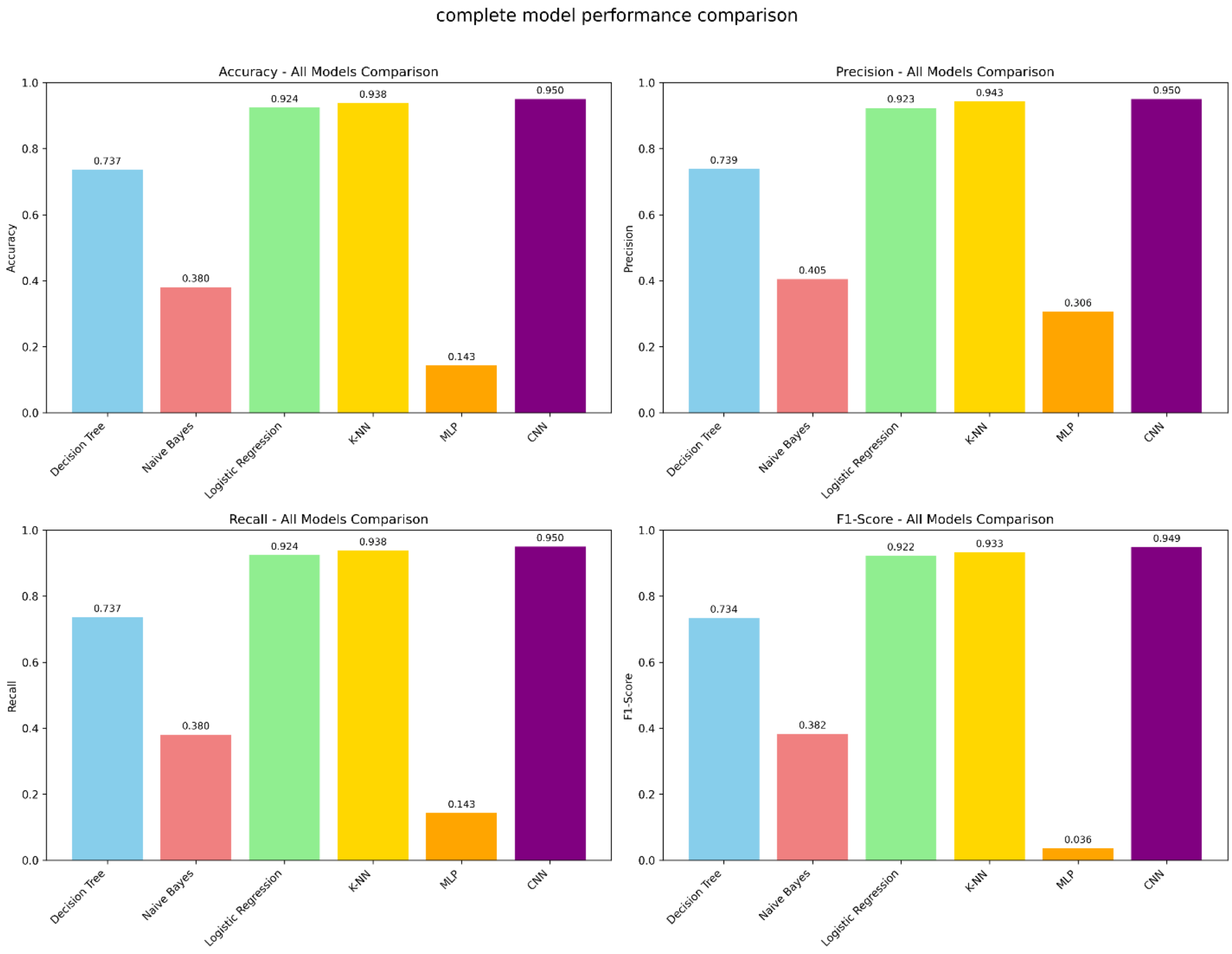


FIGURE 7: Complete Model Performance Comparison - showing all models (baseline + neural networks) across all metrics

6.3 Detailed CNN Performance Analysis

The confusion matrix for our cnn shows excellent performance across all seven classes. Dermatofibroma achieved perfect classification (100% recall) while Melanocytic nevi and Basal cell carcinoma showed best performance with over 99% accuracy. The most challenging class proved to be Actinic keratoses with 78.5% recall likely due to its visual similarity to other keratosis-like lesions.

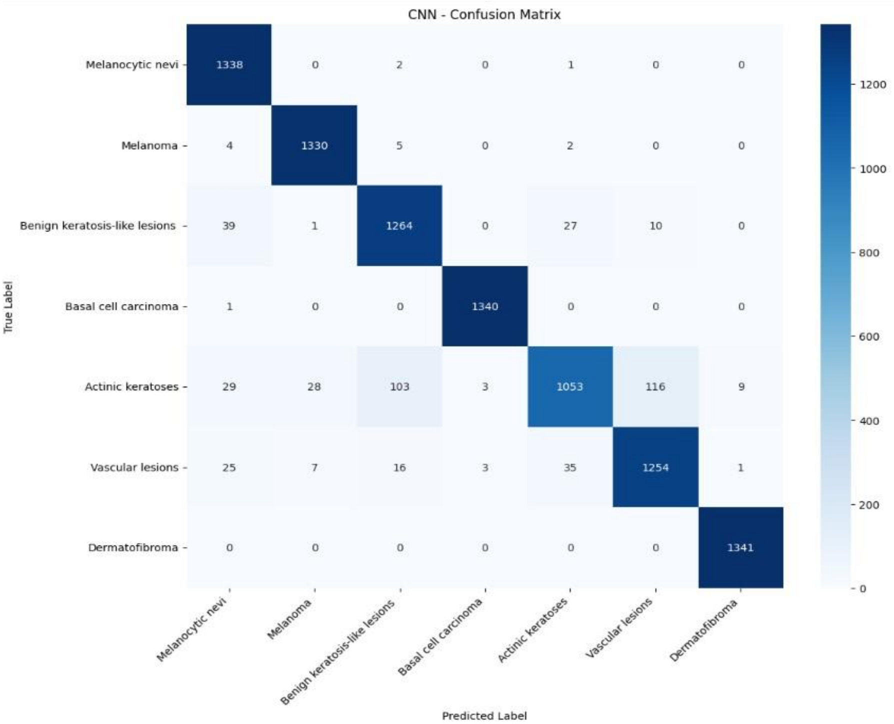


FIGURE 8: CNN Confusion Matrix - detailed heatmap showing classification results for all seven classes

Per class analysis shows that our cnn maintains high precision across all categories with values ranging from 90.9% to 99.6%. This balanced performance shows that the model has learned different features for each lesion type rather than depending upon dataset biases. The detailed classification report shows similar performance across all skin lesion categories while validating our balanced training approach.

7. GUI Implementation

To shows the practical applicability of our trained model, we developed a user friendly web application using streamlit. This implementation matches the bonus requirement for gui development while providing value for potential end users.

Our streamlit application features an interface where users can upload dermoscopic images for classification. The system automatically preprocesses uploaded images to match our training specifications by applies our best selected trained cnn model and presents results in an clear format. The interface displays the predicted class along with confidence scores and probability distributions amoung all seven categories.

The application includes suitable medical disclaimers highlighting that the tool is planned for educational purposes and should not replace professional medical diagnosis. This responsible approach acknowledges the limitations of automated systems in clinical decision making while showing the potential for AI assisted diagnostics.

8. Discussion and Analysis

Our results shows clear performance hierarchies among different algorithmic approaches to skin lesion classification. Machine learning methods when applied to flattened image data achieve good but limited performance. The success of knn shows that balanced datasets create important similarity structures in high dimensional spaces, while logistic regression strong performance shows important linear separability among classes.

The performance difference between our MLP and CNN highlights the importance of architectural choices in deep learning applications. While both networks have similar parameter counts, the CNN inductive biases toward spatial relationships and translation prove important for image classification tasks.

The CNN good performance across all metrics validates the effectiveness of our preprocessing pipeline and data balancing strategy. The model ability to achieve over 95% accuracy while maintaining balanced performance across classes shows good learning of discriminative features.

Our implementation successfully refers the practical requirements of real world deployment through the streamlit interface. This gui component transforms our research prototype into a potentially useful tool for educational purposes and shows the feasibility of automated skin lesion classification systems.

9. Challenges and Limitations

Several challenges appeared during our project implementation. The initial class imbalance required careful handling through smote though this synthetic augmentation may not fully capture the complexity of real world data distributions. The computational requirements for training neural networks on high resolution images careful architecture design to balance performance with resource constraints.

Our MLP poor performance highlights the fundamental limitations of fully connected architectures for image classification tasks. The curse of dimensionality and loss of spatial information when flattening images proved impossible despite regularization techniques.

The dataset itself while complete may not fully represent the variety of skin lesions encountered in clinical practice. Geographic and demographic biases in the HAM10000 dataset could limit generalizability to wider populations.

10. Future Work and Improvements

Several ways exist for extending and improving our work. Advanced data augmentation techniques such as rotation, scaling and color jittering could increase model robustness without needed additional labeled data. Transfer learning from pre trained models like resnet or efficientNet could improve performance while reducing training time.

The following strategies could significantly improve our system capabilities:

- **Model Architecture Improvements:** Implementation of attention mechanisms, residual connections and more complex cnn architectures could further boost classification accuracy
- **Data Enhancement:** Integration of additional datasets and advanced augmentation techniques could improve model generalization across various patient populations
- **Clinical Integration:** Development of uncertainty quantification methods and integration with electronic health records could help real-world deployment

11. Conclusion

Our complete study shows the effectiveness of machine learning approaches for skin lesion classification. The convolutional neural network shows as the superior approach which achieving 95.03% accuracy through valuable utilization of spatial relationships in dermoscopic images. This performance exceeds traditional machine learning baselines while maintaining balanced classification across all seven lesion types.

The successful implementation of our streamlit gui shows the practical feasibility of deploying such systems for educational and research purposes. Our work contributes to the growing body of research supporting ai assisted dermatological diagnosis while maintaining suitable caution about clinical deployment requirements.

The project successfully fulfilled all specified requirements including data exploration, baseline model implementation, neural network development and gui creation. The bonus feature implementation adds practical value while showing the real world applicability of our research results.

Our results shows that automated skin lesion classification systems can achieve clinically accuracy levels though careful validation and integration with existing medical workflows remain essential for practical deployment. This work provides a solid base for future research in AI-assisted dermatological diagnosis and shows the potential for machine learning to promote meaningfully to healthcare applications.

Work Distribution

The project work was distributed among team members based on individual strengths and collaborative requirements. Each phase required coordinated effort to make sure consistent methodology and complete of all project requirements.

Data Processing and Analysis Phase: This foundational work involved dataset loading, complete exploratory data analysis, visualization creation and implementation of the smote balancing strategy. The team member responsible for this component also handled missing data imputation and established the preprocessing pipeline that proved important for model training. This work represented approximately 25% of the total project effort and required deep understanding of both the medical domain and data science principles.

Baseline Algorithm Implementation: The development and evaluation of traditional machine learning approaches including decision tree, naive bayes, logistic regression, and knn classifiers required hyperparameter tuning and performance analysis. This component involved implementing proper cross-validation strategies, conducting statistical significance testing, and creating comprehensive evaluation frameworks. The responsible team member also developed the comparative analysis methodology that guided our algorithm selection process. This phase constituted approximately 30% of the project workload.

Deep Learning Development: The design, implementation, and optimization of both MLP and CNN architectures represented the most technically challenging aspect of our project. This work included extensive hyperparameter optimization, architecture experimentation, and comprehensive performance evaluation across multiple metrics. The team member leading this effort also implemented the training callbacks, learning rate scheduling, and regularization strategies that ensured successful model convergence. This component required approximately 35% of the total project effort and demanded expertise in both theoretical deep learning concepts and practical implementation skills.

Integration and Documentation: The creation of the streamlit application, complete report writing and final presentation preparation required careful coordination of all project components. This phase included implementing the user interface proper model deployment, conducting final testing, and preparing all documentation materials. The responsible team member also coordinated the final integration testing and prepared the demonstration materials. This work accounted for approximately 10% of the total effort but was crucial for demonstrating the practical applicability of our research outcomes.

Each team member contributed meaningfully to all phases of the project through regular collaboration sessions, peer review processes, and shared problem-solving activities. Individual expertise was leveraged strategically while maintaining collective ownership of all project outcomes and ensuring comprehensive understanding across the team.

References and Acknowledgments

We acknowledge the creators of the HAM10000 dataset for providing this important resource to the research community. The dataset complete nature and careful curation enabled our thorough investigation of machine learning approaches to skin lesion classification.

Special recognition goes to the open-source community for developing the tools and frameworks that made this project possible, including TensorFlow, Scikit-learn, and Streamlit. These resources demonstrate the power of collaborative software development in advancing research capabilities.