Stony Brook University

Analyzing Overfitting in Predictive Modelling of Medical Data

Abhishek Jha, Akshay Raghavan, Akshay Venkatesan

Abstract

The application of machine learning (ML) and statistical modeling in medical research has shown significant promise in disease outcome predictions and understanding various diagnosis. Nevertheless, a persistent challenge in ML-based models is overfitting. This project investigates the tendency of machine learning models to overfit when applied to medical datasets. We focus on identifying the conditions under which overfitting occurs by experimenting with various models and data preprocessing methods. Additionally, we evaluate the effectiveness of selected strategies to prevent overfitting. Our goal is to provide insights and recommendations that can guide practitioners in the development of more robust ML models.

Introduction

Predictive modeling plays a vital role in medical research, harnessing the capabilities of machine learning (ML) and statistics to unveil patterns and provide predictive insights. This approach has revolutionized our ability to extract valuable information from complex medical datasets. However, model overfitting has emerged as a pervasive and concerning issue in health research, carrying significant implications for clinical decision-making. Overfitting transpires when a model, in its endeavor to minimize errors on the training dataset, mistakenly encompasses extraneous noise rather than faithfully characterizing the underlying data distribution. Consequently, this phenomenon engenders suboptimal performance when the model is applied to novel, previously unseen data. Certain instances exemplify this issue, wherein models with high non-linearity displayed excellent in-sample performance but faltered drastically out-of-sample. Overfitting is a major risk in medical research, and this project seeks to comprehensively examine its causes prevalence, and prevention to mitigate its risk.

The Challenge of Overfitting

Overfitting in machine learning, especially in the domain of medical research, poses a significant threat to the reliability and applicability of predictive models. It occurs when a model adapts too closely to the training data, incorporating noise as part of the signal, which diminishes its ability to generalize to new data.

Scope of the Project

This project investigates the tendency of machine learning models to overfit when applied to medical datasets. Our exploration involves a detailed examination of various machine learning models, emphasizing the conditions that lead to overfitting and evaluating methods to mitigate this issue.

Methodology

Our methodology revolves around two central hypotheses:

1. **Data-Induced Overfitting:** We hypothesize that intrinsic noise, sampling biases, and data imbalance contribute to overfitting. This is explored using statistical methods and selective training on datasets like PIMA Indian Diabetes and Skin Cancer MNIST. We aim to ascertain the correlation between these data characteristics and overfitting tendencies.
2. **Model-Induced Overfitting:** We focus on the premise that model complexity and pre-training significantly impact overfitting. Different configurations of ANNs and CNNs, including their depth and parameter count, will be evaluated. We will compare pretrained networks with those trained from scratch, particularly for image-based datasets, to assess their overfitting behaviors.

This approach aims to dissect the multifaceted nature of overfitting in medical ML models, providing insights into both data and model-related factors.

Stony Brook University

Literature Review

Background

Machine learning (ML) is pivotal in scientific advancements but is often marred by overfitting—models perform well on training data but poorly on unseen data. This review examines scholarly articles addressing overfitting in various medical research contexts, highlighting the delicate balance that needs to be maintained between model complexity and data integrity. The challenges of overfitting span various aspects of medical research, from immunology to epidemiology, underscoring the need for robust model evaluation and validation techniques.

Overfitting in Immunological Applications: An Exploration of Challenges and Solutions

Gygi et al. (2023) delve into the issue of overfitting in immunological data, particularly in the context of high-dimensional datasets common in vaccine development. The authors suggest that regularization techniques, such as L1 and L2 penalties, are effective in reducing overfitting. They emphasize the importance of cross-validation in building complex yet generalizable models, which is crucial for accurate vaccine development and immunological predictions. Their findings advocate for a careful balance between model accuracy and generalization, especially in fields where precision is paramount.

Navigating Nonlinearity and Overfitting with COVID-19 Data

Peng and Nagata (2020) discuss the intricacies of dealing with nonlinear COVID-19 data and the associated risk of overfitting. Their study advocates for dimensionality reduction strategies to manage model complexity and the use of information criteria for optimal model selection. They also highlight the effectiveness of ensemble methods in stabilizing predictions, a vital factor in accurately modeling pandemic trends and making reliable COVID-19 predictions. The study underscores the importance of adaptive models in rapidly evolving scenarios like pandemics, where data characteristics can change swiftly.

Ensemble of Adapted Convolutional Neural Networks for Classifying Colon Histopathological Images

In their study on using CNNs for colon cancer image classification, Albashish et al. (2022) propose an ensemble learning approach. By fine-tuning models like DenseNet121 and MobileNetV2 and introducing layers that manage overfitting, such as dense and dropout layers, they demonstrate improved accuracy in medical image analysis. This approach signifies a substantial advancement in employing CNNs for accurate classification of histopathological images in medical diagnostics, offering insights into how ensemble methods can enhance the reliability and robustness of deep learning models in medical imaging.

Responsible ML in Medicine: Navigating Regulatory Conformity

Petersen et al. (2022) approach overfitting from the perspective of regulatory compliance in medical ML. They discuss the use of transparent and explainable AI models that conform to regulatory standards. Their recommendations include expanding training sets with synthetic data and employing robust validation methods, like temporal and external validation, to ensure that medical ML models are reliable and generalizable across diverse patient groups, aligning with regulatory expectations. This highlights the growing need for ML models in medicine to be not only accurate but also understandable and transparent, ensuring patient safety and adherence to medical standards.

Predictive Modeling in Medicine: Balancing Complexity and Performance

Toma and Wei's (2023) article in the Encyclopedia provides a comprehensive overview of predictive modeling in medicine. They discuss various modeling techniques, including machine learning, and their applications in medical diagnosis and prognosis. The paper sheds light on the critical balance between model complexity and predictive performance, a key concern when considering overfitting.

Stony Brook University

Dataset and Experiment Methodology

Data Acquisition and Description

PIMA Indian Diabetes Dataset: Sourced from Kaggle, this dataset is a compilation of medical diagnostic measurements for diabetes. It comprises 768 instances with 8 features including Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age. The dataset is selected for its representation of key clinical data points and its propensity to induce overfitting due to its relatively small size and high feature count, presenting a challenge in developing generalizable models.

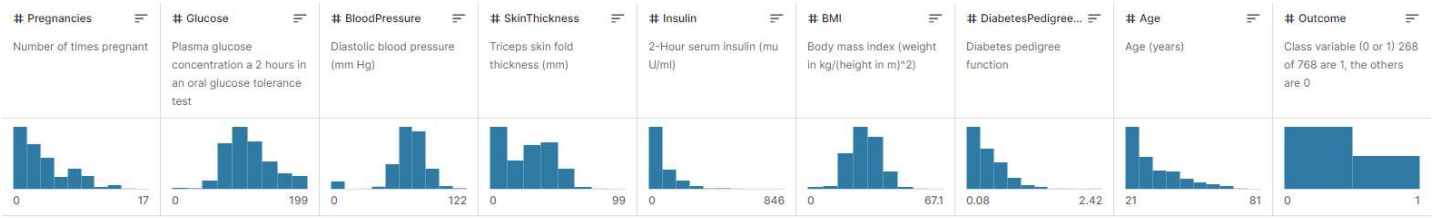


Figure 1:PIMA Diabetes Data Structure

Skin Cancer MNIST: HAM10000 Dataset: Also available on Kaggle, this dataset consists of 10,015 dermoscopic images across different skin lesion types. Each image is annotated with diagnostic information, offering a rich resource for image-based classification tasks. The selection is based on the dataset's visual complexity and its tendency to cause overfitting in image recognition models, particularly due to the high dimensionality of the image data.

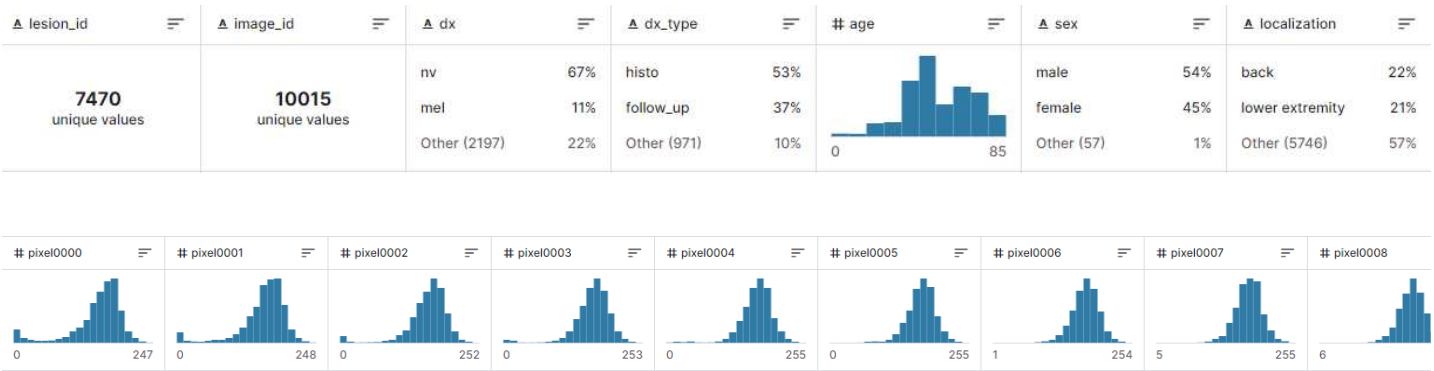


Figure 2:HAM1000 Metadata and Image Data Structure

Data-Induced Overfitting

Presence of Noise: Hypothesis that intrinsic noise in a dataset leads to overfitting, assessed by using statistical methods to identify noise and its correlation with model performance.

Sampling Bias: Hypothesis that biased sample training causes overfitting, tested through training models on specific subsets of the PIMA dataset and analyzing generalization.

Imbalance in Data: Hypothesis that imbalanced datasets lead to model overfitting towards the majority class, examined using the Skin Cancer MNIST dataset to evaluate model accuracy on minority classes.

Curse of Dimensionality: Hypothesis that high data dimensionality relative to sample size leads to overfitting, tested by adjusting the image resolution of the Skin Cancer MNIST dataset and analyzing the impact on overfitting.

Model-Induced Overfitting

Too Many Parameters: Hypothesis that models with a high number of parameters are more susceptible to overfitting, explored by testing various ANNs on the PIMA dataset and CNNs on the Skin Cancer MNIST dataset.

Pretrained vs Training from Scratch: Hypothesis that pretrained networks exhibit less overfitting compared to those trained from scratch, especially in image-based datasets, evaluated by comparing models on the Skin Cancer MNIST dataset.

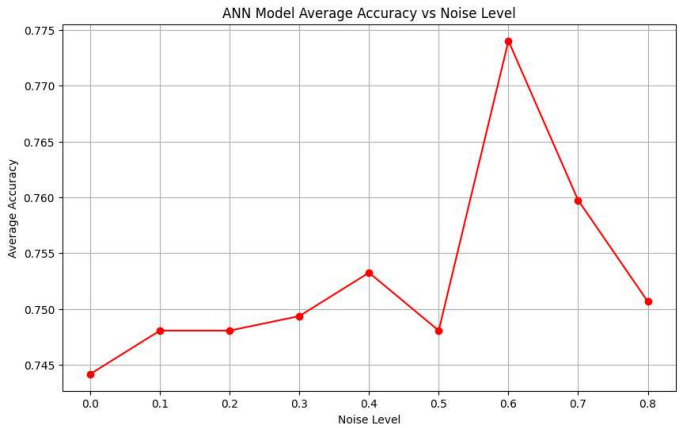
Hypothesis - 1: Presence of Noise in Data

Objective

The primary objective of this study is to investigate the impact of intrinsic noise within the PIMA Indian Diabetes Dataset on the overfitting of machine learning models, particularly artificial neural networks (ANNs). This dataset, sourced from Kaggle, includes medical diagnostic measurements for diabetes, making it prone to overfitting due to its relatively small size. We hypothesize that as the noise level increases, the generalization ability of the ANN decreases, leading to overfitting.

Methodology

We employed the PIMA Indian Diabetes Dataset, comprising 768 instances with 8 clinical features, to train an ANN. The dataset was systematically modified by introducing Gaussian noise at various levels, ranging from 0% to 80%. The ANN model was trained on these progressively noised datasets, and both training and test accuracies were evaluated to observe the effects of noise on the model's performance. This approach allows us to determine the efficacy of noise reduction techniques compared to L1 regularization in the development of generalizable models.



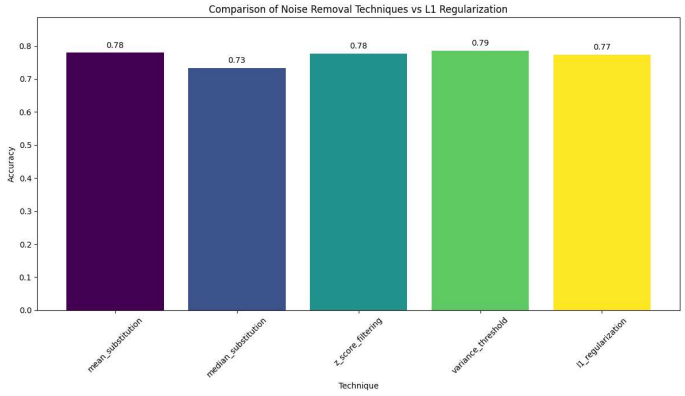
Observations

- **Noise Sensitivity:** The ANN model demonstrated varying sensitivity to different levels of noise. At some noise levels, there was a noticeable drop in accuracy, indicating that the model was indeed impacted by the noise in the data.
- **Performance Fluctuations:** The performance of the ANN fluctuated with increasing noise levels. This suggests that the model's ability to generalize is affected by the presence of noise in the data, leading to less consistent predictions.
- **Overfitting Indication:** While not directly measured, the divergence in performance with increasing noise levels could indicate a propensity for overfitting. The model may be capturing the noise as a feature, which diminishes its predictive power on the noise-free test set.
- **Non-linear Relationship:** The relationship between noise level and model accuracy was non-linear, showing that the impact of noise is not always directly proportional to its amount. Some noise levels led to a decrease in accuracy, while others did not have a significant impact or showed an increase.
- **Hypothesis Validation:** The data supports the hypothesis that intrinsic noise in the dataset can lead to overfitting in machine learning models. The ANN's varying accuracy at different noise intensities demonstrates that noise can affect the model's learning process.
- **Data Quality Importance:** The experiment underscores the importance of data quality in training machine learning models. Ensuring data cleanliness and reducing noise are critical steps in preparing data for model training.

Mitigation Techniques

- **Mean Substitution:** Replaces noisy data points with the mean value of the feature across the dataset. This method can smooth out extreme values in data but may not be suitable when the noise is systematic or the data contains many outliers.
- **Median Substitution:** Substitutes noisy data points with the median value of the feature. It is more robust than mean substitution, especially in the presence of outliers, as the median is less affected by extremely large or small values.
- **Z-Score Filtering:** Employs z-scores to identify and remove outliers based on standard deviation. This technique assumes a normal distribution of data and works best when the noise is random and can be considered as outliers.
- **Variance Threshold:** Removes features with low variance, under the assumption that features with little variance do not contribute significantly to the model's predictive power. It is particularly effective when the dataset contains features with constant or near-constant values.
- **L1 Regularization:** Also known as Lasso regularization, it adds a penalty equal to the absolute value of the magnitude of coefficients to the loss function. This technique not only helps in reducing overfitting but can also lead to sparse models where some feature weights can become exactly zero, effectively performing feature selection. L1 regularization is useful when you want to reduce the complexity of a model by penalizing the number of features

Performance Comparison: Among the evaluated techniques, the *Variance Threshold* method showed the best performance, achieving the highest accuracy. This suggests that in this specific case, removing features with low variance contributed to a more robust model by eliminating irrelevant features that could be considered as noise.



Conclusion

The results showed a non-linear relationship between noise levels and test accuracy. While some levels of noise led to a slight increase in accuracy, possibly due to the model's ability to ignore the noise, higher levels consistently resulted in decreased accuracy, supporting the hypothesis that noise can lead to overfitting. The study confirmed that intrinsic noise in training data could lead to overfitting in ANNs. The findings emphasize the importance of data preprocessing and the need for robust model evaluation metrics to ensure the generalizability of machine learning models.

Hypothesis - 2: Sampling Bias

Objective

Can bias induced to the model through choice of data increase overfitting of the model? There are various ways bias can persist, anytime during the pipeline of ML. We hypothesize to explore the impact of selection and sampling biases on the PIMA diabetes dataset.

Methodology

The dataset of concern is PIMA wherein we diagnostically predict whether a patient has diabetes. It involves information of the patient like blood pressure, bmi index, insulin levels, skin thickness, age and more. The dataset consists of 768 instances in total wherein 500 samples belong to outcome '0' and the rest to outcome '1'

	type	from_range	to_range	train_accuracy	test_accuracy
0	age	25	45	0.979769	0.714286
1	bmi	28	50	0.978622	0.727273
2	bloodpressure	60	100	0.984344	0.740260
3	baseline	0	0	0.978827	0.759740

Figure 1:Train and Test Accuracy

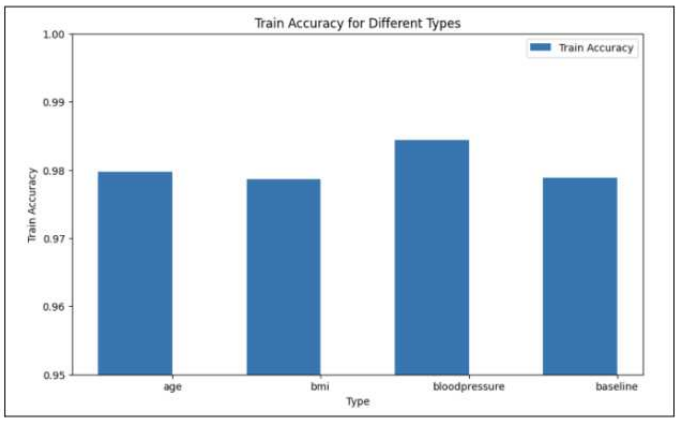


Figure 2:Train Accuracy

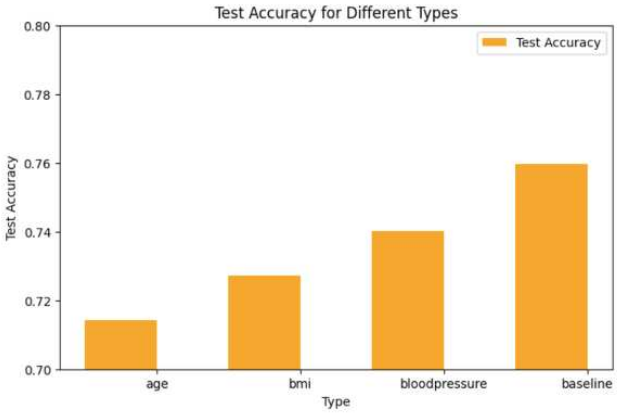


Figure 3:Test Accuracy

Observations

Sampling by Age, BMI, and Blood Pressure:

- Variations in training and test accuracies upon altering age range highlight sampling bias. This phenomenon indicates the model's skewed proficiency towards specific age groups, leading to challenges in predicting outcomes for ages less represented in the training data. This trend underscores the importance of diverse age representation for balanced model performance.
- The observed accuracy fluctuations with age range modifications emphasize the model's dependency on the training data's age distribution. Narrow age ranges in training lead to reduced accuracy in testing across broader age groups, pointing to the necessity for incorporating a wide age spectrum in the training set for better generalization capabilities.

Stratify vs Random Sampler:

- Stratified sampling, designed to mirror the original dataset's class distribution in training and testing sets, may inadvertently lead to an overemphasis on the majority class. This scenario is particularly evident when excluding specific value ranges, resulting in a skewed class distribution and biased model predictions.
- The random sampler, in contrast, ensures a more eclectic representation in the training set by not strictly adhering to class proportions. This diversity can be crucial in training more robust models, highlighting the importance of selecting a sampling strategy that aligns with the dataset's unique characteristics to avoid biases and enhance model reliability.

Model Performance and Generalization:

- Test accuracies of 0.75, 0.74, and 0.73 on biased subsets, as opposed to a consistent 0.75 on the full dataset, point to the model's challenges in generalizing across diverse data distributions. This subtle discrepancy in accuracy between biased subsets and the full dataset suggests the model's limitations in adapting to broader dataset variations.
- To address these generalization challenges, it's essential to rigorously evaluate the training set for potential biases and implement strategies like collecting diverse data and using stratified sampling. Additionally, analyzing the model's performance across various underrepresented subsets provides critical insights into enhancing its predictive accuracy and overall robustness.

Mitigation Techniques

- Conduct careful experimentation and deploy strategies like random sampling and stratified sampling.
- Implement standards and clear objectives during data collection and experimentation.
- Make efforts to ensure the dataset is representative of the broader population.
- Use Cross-Validation to identify and mitigate the impact of sampling bias on model generalization.
- Perform Sensitivity Analysis to evaluate how different assumptions about the sampling process impact study conclusions.
- **Snowball Sampling:** Start with a small group of participants and expand using their networks, especially useful for hard-to-reach populations.
- **Weighting:** Apply weights to adjust for overrepresented or underrepresented groups in the sample, correcting for imbalance and improving representativeness.

Conclusion

Sampling bias poses a real threat to research results. However, the consequences of this type of bias extend beyond skewed results. Sampling bias can be inadvertently induced in such small datasets hampering the model's generalizability.

Hypothesis - 3: Imbalance in Data

Objective

Imbalances in a dataset can occur when there exists a significant skewness in the distribution of the dataset. The generalizability of the model can be sensitive in their performance due to these class imbalances. We set to evaluate this hypothesis on the infamous HAM1000 "Human Against Machine with 10000 training images" dataset and propose mitigation strategies.

Methodology

The dataset adopted is diagnosis of pigmented skin lesions. It's a multi-modal dataset with images of the lesions, characterized with information like sex of the patient, localization/geographical location of the lesion on the patient, etc. It's a classification problem with severn classes consisting of 10,015 dermatoscopic images. This dataset was inspired to bridge the gap in small size and lack of diversity of available dataset of dermatoscopic images. However, the imbalance is quite significantly dominated by Melanocytic nevi class 67% of the time.

Being a multi-modal data, the strategies proposed are particular towards computer vision techniques that can be employed on images. To test the hypothesis and propose mitigation strategies, a vanilla imagenet pre-trained ResNet-50 architecture was used as baseline. The class imbalance had to be effectively mitigated through techniques like downsampling, upsampling and more. The images were augmented through randomized rotation, horizontal flips, vertical flips. shifts in the lateral and vertical axeses. The minority classes were oversampled adopting the mentioned techniques to increase their frequency to match the majority class. Undersampling is a method to reduce the majority class to match the minority class frequency. This was however not adopted as to avoid the loss of data for training.

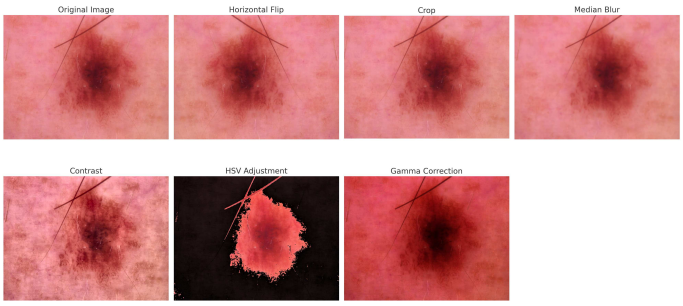


Figure 1:Process of Image Augmentation

Observations

Resampling Techniques and Test Metrics vs Dataset: The model's performance slightly decreased despite a 7x increase in the dataset for training, indicating better generalization and more representations learned of the minority class.

Oversampling vs Baseline Model: The baseline model's performance plateaued, while the oversampling model showed a gradual increase in its metrics, indicating improved handling of class imbalances.

Validation and Test Cutouts: Utilization of validation cutouts of images increased the model's performance on unseen data, bridging the gap between validation and test errors.

Metrics of Choice: In a multi-class classification of an imbalanced dataset, metrics like f1 score and ROC AUC are preferred over accuracy. This approach mitigates the model's tendency to predict the majority class and provides better insight into its performance in predicting minority classes.

Shuffle vs No Shuffle: A notable observation during model training was high losses with classifications skewed towards the majority class. This was attributed to inadequate shuffling in train and validation splits. Implementing proper shuffling strategies led to faster convergence.

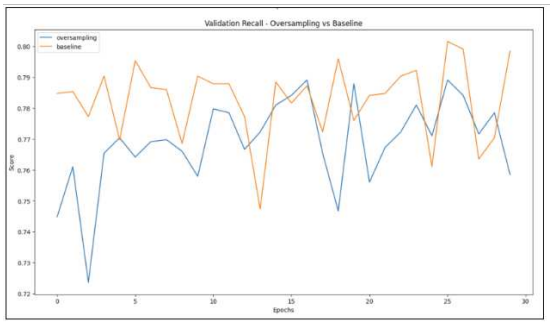


Figure 2:Recall Analysis

Mitigation Techniques

Catch them with your metric!

- Move beyond accuracy; consider Precision, Recall, F1 Score, MCC, AUC for a holistic view.

Handling Chaos in your Classes!

- **Undersampling Majority:** Techniques include Controlled, Cleaning, Prototype, Random Sampler, Near-Miss, Tomek Links, Neighbourhood Cleaning Rule, One Sided Selection, Cluster Centroids, Edited Nearest Neighbours, All K-Nearest Neighbors, Condensed Nearest Neighbour, Instance Hardness Threshold.
- **Oversampling Minority:** Approaches like Random Sampler, ProWSyn, AdaSyn, Bootstrapping, SMOTE, Modified-SMOTE, Polynom-fit-SMOTE, SMOTE-IPF, Borderline-SMOTE, SVM-SMOTE.

Validate Before It's Too Late

- Employ K-fold Cross Validation. Note the importance of cross-validation before over-sampling to avoid overfitting to artificial bootstrapping results.

Ensembles Assemble - The ML Avengers

- Techniques include Bagging, SMOTEBagging, UnderBagging, OverBagging, Boosting, RUSBoost, EasyEnsemble, Balanced Random Forests.

Sometimes, Be Partial?

- Implement Weight Balancing Loss, biasing towards rare or difficult samples. Techniques include inverse-frequency weighting (INV) and Focal Loss.

Know your Domain - Thoroughly!

- Consider case-control studies for deeper insights.

```
{'nv': 4301, 'bcc': 4176, 'mel': 4320, 'vasc': 3761, 'akiec': 4282, 'bkl': 4230, 'df': 3780}
```

Figure 3:Classwise Distribution after balancing

Conclusion

When dealing with significantly imbalanced datasets, there's a risk of models overfitting towards the predominant class. To address this, we implemented a strategy of oversampling the underrepresented classes to equalize label representation. Despite this approach, our experiment showed no notable enhancement in performance on the oversampled dataset compared to the original. This lack of improvement might be specific to our particular problem, suggesting that outcomes could vary with different datasets and model combinations.

Hypothesis - 4: Curse of Dimensionality

Objective

The primary aim of this study is to investigate the impact of increasing feature-space dimensionality on the overfitting tendencies in machine learning models and focuses on the Skin Cancer MNIST: HAM10000 dataset, notable for its high-dimensional imagery and limited sample size. Contrary to the usual approach of reducing features to mitigate overfitting, this study posits that an increase in the ratio of features to samples will exacerbate overfitting issues. We intend to test this hypothesis by progressively enhancing image resolution, thereby increasing the feature count, to evaluate whether this increment leads to more pronounced overfitting in complex image-based tasks.

Methodology

In this study, we employed the Skin Cancer MNIST: HAM10000 dataset from Kaggle, which comprises 10,015 dermoscopic images across various lesion types. Each image, rich in diagnostic detail, presents a substantial challenge for image classification models due to its visual complexity. To examine the effects of image resolution on model overfitting, we trained a Convolutional Neural Network (CNN) on this dataset at five different resolutions: 64×64, 128×128, 256×256, 512×512, and 1024×1024. Model performance was gauged using the accuracy and generalization error at each resolution.

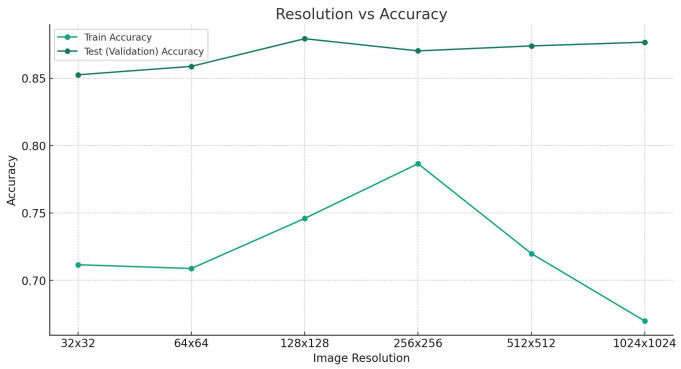


Figure 1:Resolution vs Accuracy

Figure 2:Resolution vs Generalization Error

Observations

- **Diminishing Returns:** There is an evident point of diminishing returns regarding resolution enhancement, where further increasing the resolution does not yield better performance and may indeed harm the model's generalization capabilities.
- **Resolution and Accuracy:** As image resolution increases from 32×32 to 256×256, there is a general trend of increased accuracy in both training and validation phases, suggesting that higher resolution images contain more useful features for the model. However, beyond 256×256, accuracy plateaus or slightly decreases, indicating a potential overfitting threshold.
- **Generalization Error:** The generalization error first decreases as resolution improves, reaching a minimum at 256×256, which could be considered the optimal resolution for balancing detail with model performance. Beyond this point, the error increases sharply, highlighting a decrease in the model's ability to generalize.
- **Overfitting at High Resolutions:** The sharp rise in generalization error at resolutions higher than 256×256 indicates that the model may be overfitting to the training data, capturing noise rather than relevant features.
- **Optimal Resolution Hypothesis:** These observations support the hypothesis that there is an optimal image resolution for training the model that provides a sufficient level of detail while avoiding the introduction of unnecessary noise or complexity that could lead to overfitting.

Mitigation Techniques

For image-based datasets, the following dimensionality reduction techniques can be particularly effective:

- **Principal Component Analysis (PCA):** Useful in transforming high-dimensional image data into a lower-dimensional form by focusing on the principal components that capture the maximum variance.
- **Autoencoders:** Employ neural networks to learn a compressed representation of images, aiding in reducing dimensions while retaining key features.
- **t-Distributed Stochastic Neighbor Embedding (t-SNE):** Primarily used for high-dimensional data visualization, including image datasets, by converting similarities into joint probabilities in a lower-dimensional space.
- **Feature Selection:** In image processing, this can involve techniques like edge detection, color space reduction, or texture analysis to select relevant features.

For textual data, dimensionality reduction can be approached through:

- **Latent Semantic Analysis (LSA):** Similar to PCA but specifically adapted for text data, it reduces dimensions by identifying patterns in relationships between terms and concepts.
- **Word Embeddings:** Techniques like Word2Vec or GloVe provide a dense representation of words in a lower-dimensional space.
- **Feature Hashing:** Converts textual features into a fixed size lower-dimensional space using hashing.
- **Feature Selection:** In text data, this might involve selecting specific keywords, phrases, or syntactic patterns that are most relevant for the task.

Both sections aim at reducing the complexity of the model to mitigate overfitting in their respective data types.

Conclusion

Our findings indicate that there is a non-linear relationship between image resolution and model accuracy in the context of skin lesion classification. An optimal resolution exists (256×256) at which the model achieves balanced accuracy and generalization, suggesting that beyond this point, additional detail does not equate to improved model performance and may in fact contribute to overfitting. These results affirm the importance of feature-space dimensionality in model training and underscore the necessity of selecting an appropriate image resolution to maximize model efficacy while minimizing the risk of overfitting.

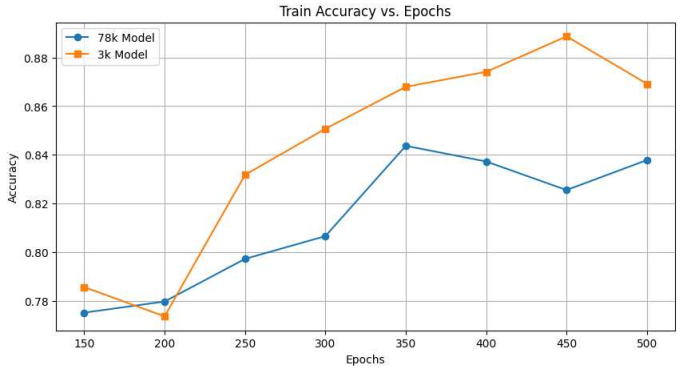
Hypothesis - 5: Too many model parameters

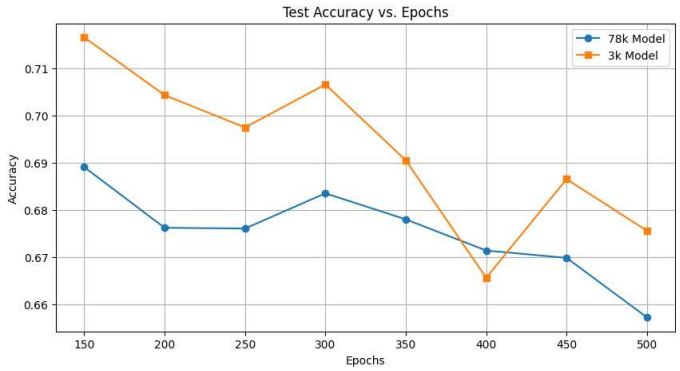
Objective

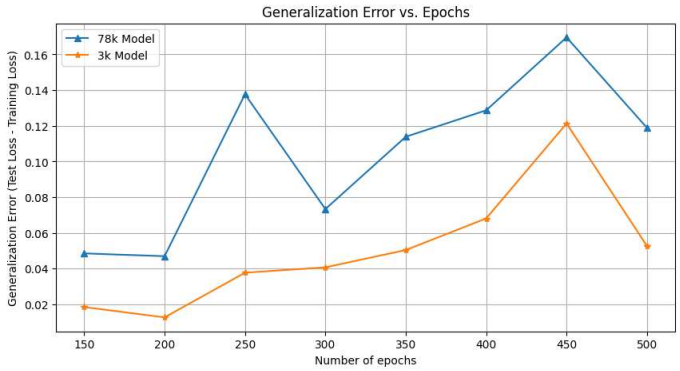
It is often found that models that are overly complex, characterized by having too many parameters, tend to be more prone to overfitting. This insight led us to hypothesize that there is a direct correlation between the number of parameters in a model and its likelihood to overfit on any given dataset. For this experiment, we picked PIMA Indian Diabetes dataset and the model of our choice was an Artificial Neural Network.

Methodology

Initially, we trained an ANN consisting of approximately 3,000 parameters, running it through 200 to 500 epochs. Following this, we escalated the complexity of the model by increasing the number of parameters to approximately 78,000, subjecting it to an equivalent number of epochs. Throughout this process, we meticulously observed and recorded the trends in both training and testing losses, as well as the accuracy of each model across the varying epochs, to see if there was a significant difference in the performance of both models.





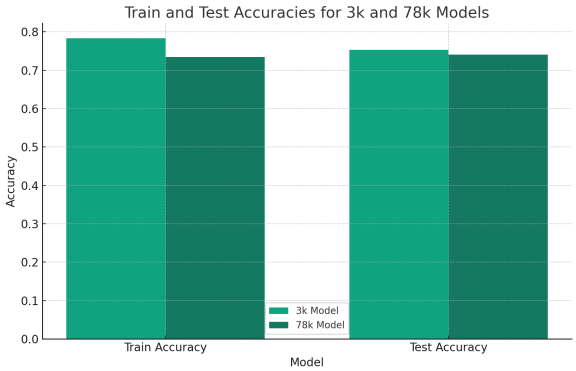


Observations

- **Test Accuracy vs. Epochs:** The 78k parameters model shows greater fluctuations in test accuracy compared to the 3k parameters model, suggesting less stability in its generalization capability. The 3k parameters model demonstrates a more consistent test accuracy, which could imply better generalization than the 78k model.
- **Generalization Error vs. Epochs:** The generalization error for the 78k model is higher and more variable across epochs, signifying a discrepancy between its performance on training and test data, a hallmark of overfitting. The 3k model maintains a lower and more stable generalization error, suggesting it is not overfitting as much and is likely a better fit for the given data.
- **Training Performance:** The 78k model's training accuracy plateaus earlier and remains consistently high, suggesting that it may be capturing noise and nuances specific to the training set rather than underlying patterns applicable to the general data. The 3k model's training accuracy shows a steady, less steep increase, which typically correlates with a model learning the underlying structure of the data rather than memorizing it.

Mitigation Techniques

- **Early Stopping:** We mitigated overfitting by capping the training at 70 epochs. This early stopping prevented the model from internalizing training set noise, thereby maintaining its generalization capabilities.
- **Adhering to a simpler model:** Overfitting can also be addressed by opting for the less complex 3k model. Its reduced capacity to over-learn minute training details helped in achieving a more generalized performance on unseen data. Both of these points can be visualized by the attached comparison bar plot.



Conclusion

In conclusion, our analysis revealed that models with higher complexity, such as the 78k parameter model, are more prone to overfitting as evidenced by their fluctuating test accuracy and higher generalization error. By limiting training epochs and employing simpler models, we effectively enhanced model generalization. The 3k model, with its lower parameter count, demonstrated more stable performance, underscoring the importance of model simplicity and training restraint in preventing overfitting and ensuring robust predictive performance.

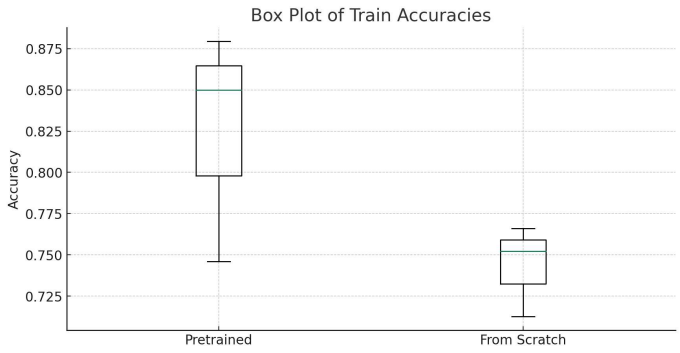
Hypothesis - 6: Pretrained vs From Scratch models

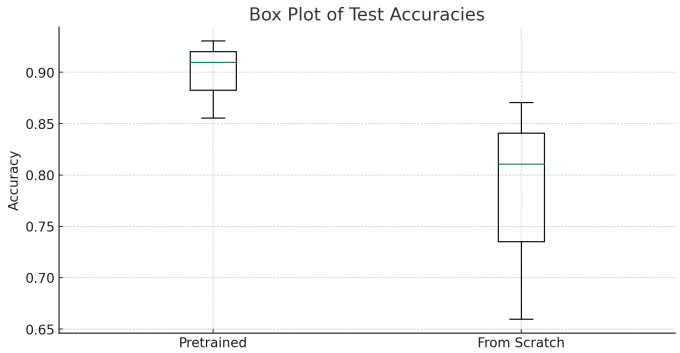
Objective

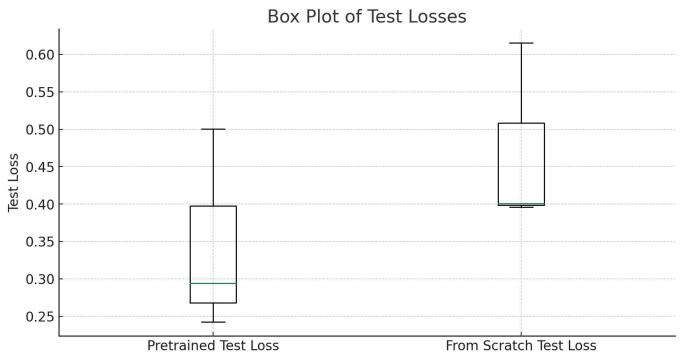
We aim to assess the hypothesis that pretrained neural networks show reduced overfitting and superior performance on image-based datasets, as opposed to models trained from scratch. This is predicated on the observation that pretrained models often benefit from transfer learning, which potentially equips them with a more generalized understanding of features, leading to enhanced predictive accuracy in diverse imaging contexts.

Methodology

Our study involves training ResNet34, VGG11, and DenseNet121 on the SKIN CANCER MNIST dataset, both with pretrained weights and from scratch, for 25 epochs. This dataset, predominantly skewed with 67% of data in one class, serves as a challenging benchmark for overfitting. The goal is to compare the performance of pretrained models against those trained from scratch, evaluating whether pretrained networks demonstrate better resistance to overfitting and improved overall performance, thus supporting our hypothesis in the context of image-based datasets.





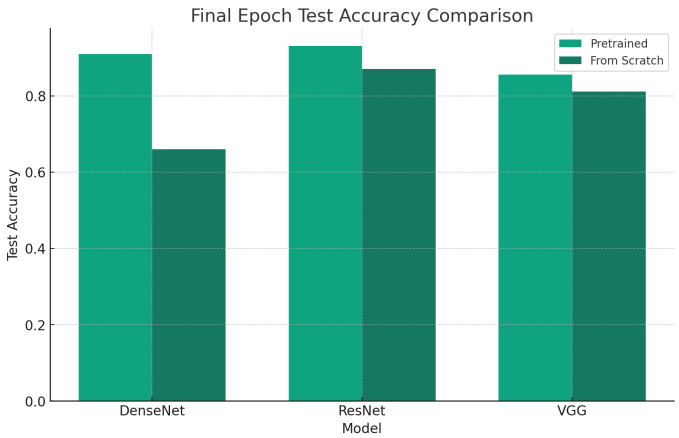


Observations

- **Train Accuracies:** The pretrained models generally exhibit higher median train accuracies compared to the models trained from scratch, indicating better learning from the training dataset. The spread (interquartile range) of train accuracies is narrower for pretrained models, suggesting more consistent performance across different architectures.
- **Test Accuracies** Pretrained models also show higher median test accuracies than their from-scratch counterparts. This implies better generalization to unseen data. There is less variability in the test accuracies of pretrained models, as indicated by the tighter interquartile ranges, pointing to more reliable and stable performance on new data.
- **Comparative Consistency** The from-scratch models show a wider spread in both train and test accuracies, indicating a less consistent performance across different model architectures. The pretrained models' box plots are more compact, suggesting that the benefit of pretraining is more uniformly experienced across different network architectures.
- **Median Test Loss** The median test loss for pretrained models appears to be lower than that of the from-scratch models, indicating that on average, pretrained models tend to have better performance (lower loss) on unseen data.

Mitigation Techniques

- **Prefer pretrained models:** The attached plot unequivocally shows that pretrained models are less prone to overfitting on training data, as reflected by their superior test accuracies in all three model types: DenseNet, ResNet, and VGG. This phenomenon becomes increasingly pronounced with more complex networks. For instance, the improvement in test accuracy for DenseNet's pretrained version compared to its from-scratch counterpart is significantly larger than the improvements observed in the other two models, indicating a stronger benefit of pretraining in more complex network architectures.



Conclusion

The investigation into pretrained versus from-scratch models underscores the effectiveness of pretrained models in mitigating overfitting. This enhanced robustness is largely due to the transfer of features learned from extensive, diverse datasets, which these models were originally trained on. Leveraging pretrained models taps into their vast repository of prior learning, leading to improved generalization on new tasks and datasets. Consequently, the adoption of pretrained models emerges as a potent strategy, ensuring more reliable and universally applicable results in various practical scenarios.

Stony Brook University

Recommendations to practitioners

H1: Presence of Noise

Outcome: The study found that noise in training data impacts the accuracy of Artificial Neural Networks (ANNs) in a non-linear way. Initially, moderate noise levels improved accuracy, likely because the model learned to overlook the noise. However, as noise levels increased, accuracy decreased, indicating overfitting. This confirms the hypothesis that excessive noise leads to overfitting in ANNs.

Recommendation: In light of the ANN's sensitivity to varying noise levels and its fluctuating performance, it is recommended that practitioners employ adaptive noise management and dynamic model tuning. This involves using noise filtering techniques tailored to different noise intensities and adjusting model parameters in response to noise variations. Also, given the non-linear relationship between noise levels and accuracy, non-linear noise compensation strategies should be employed. Robust validation mechanisms, such as cross-validation with noise-augmented data, are essential for assessing model generalizability. Continuous monitoring and adjustment of the model in response to real-world noise scenarios will also be crucial for maintaining optimal performance.

H2: Sampling Bias

Outcome: The experiment showed that the model's accuracy on biased subsets of the training dataset was marginally lower compared to its performance on the entire dataset, suggesting challenges in generalizing across diverse data distributions. This indicates the model's limited adaptability to varied data, emphasizing the need to scrutinize and diversify the training dataset for representativeness. Employing strategies like inclusive data collection and stratified sampling can help reduce bias. Regularly evaluating the model's performance on diverse subsets is also crucial to enhancing its generalization capabilities and overall predictive accuracy.

Recommendations: Based on the obtained observations, it is recommended that practitioners pay close attention to the diversity of their training datasets - ensure that the training data encompasses a broad spectrum of age groups to better represent the diversity of the broader population. This will help the model perform more accurately across varied categories, not just those predominant in the training set. Practitioners should carefully assess their sampling strategy to ensure it aligns with the goal of creating a balanced and representative dataset, thereby enhancing the model's ability to generalize effectively across diverse groups and categories.

H3: Imbalance in Data

Outcome: When dealing with significantly imbalanced datasets, there's a risk of models overfitting towards the predominant class. To address this, we implemented a strategy of oversampling the underrepresented classes to equalize label representation. Despite this approach, our experiment showed no notable enhancement in performance on the oversampled dataset compared to the original. This lack of improvement might be specific to our particular problem, suggesting that outcomes could vary with different datasets and model combinations.

Recommendations: Given these findings, it is recommended that practitioners approach the issue of dataset imbalance with a degree of caution. While oversampling the minority class is a common strategy, its effectiveness can vary significantly depending on the specific problem and dataset-model combination. Therefore, it's advisable to first analyze the unique characteristics of your dataset and the nature of the imbalance. Experiment with a range of techniques, such as different oversampling methods, synthetic data generation, or even advanced ensemble methods, to see what works best in your specific context. It's also important to remain open to the possibility that in some cases, traditional imbalance mitigation strategies might not yield significant improvements, and exploring alternative, perhaps more problem-specific solutions could be more effective.

H4: Curse of Dimensionality

Outcome: The study reveals a complex interaction between image resolution and model precision in skin lesion classification. We identified an ideal resolution (256x256) where the model attains a harmonious balance between accuracy and generalizability. It appears that resolutions higher than this threshold do not enhance, and might even hinder, model performance, potentially leading to overfitting. This underscores the significance of feature-space dimensionality in training and highlights the critical need to choose a suitable image resolution to optimize model effectiveness and reduce the likelihood of overfitting.

Recommendations: Given these insights, it is recommended that practitioners pay close attention to the impact of data features on model training. An observed optimal resolution (256x256 in the case study) marks a balance between detail inclusion and model performance, where accuracy peaks and overfitting risks are minimized. It's crucial to recognize the point of diminishing returns in resolution enhancement. Practitioners should strive to identify this optimal resolution in their specific context. Monitoring generalization error is key to determining this balance, as it helps in identifying when higher resolutions cease to contribute to model effectiveness and start to hinder its generalizability. This approach is broadly applicable and can be adjusted based on the specific requirements and characteristics of different datasets and models.

H5: Too many model parameters

Outcome: We found that models with greater complexity, like ours with 78,000 parameters, tend to overfit, as shown by their inconsistent test accuracy and increased generalization error. We also discovered that reducing the number of training epochs and using less complex models significantly improved their ability to generalize. The model with only 3,000 parameters showed more consistent results, highlighting the value of simpler model architectures and controlled training in preventing overfitting and achieving reliable predictive accuracy.

Recommendation: It's recommended for practitioners to opt for models with lower parameter counts, like the 3k parameter model, which demonstrated more consistent test accuracy and less fluctuation, indicating better generalization capabilities. Avoid models with excessively high parameters, such as the 78k model, as they tend to show higher and more variable generalization errors. Additionally, observe the training performance closely; models that plateau early and show consistently high training accuracy might be capturing noise and specific training set nuances, leading to overfitting. Aim for a steady increase in training accuracy as such balanced approach will help in developing models that not only perform well on training data but also generalize well.

H6: Pretrained vs From Scratch models

Outcome: This study highlighted the superior ability of pretrained models to reduce overfitting. This improved robustness stems from the transfer of features learned from large and varied datasets, on which these models were initially trained. It was found that utilizing pretrained models allows access to their extensive prior learning, resulting in better generalization for new tasks and datasets. Therefore, employing pretrained models proves to be a powerful approach, yielding more dependable and broadly applicable outcomes in diverse practical applications.

Recommendation: Practitioners are advised to favor pretrained models over those trained from scratch, given their demonstrated superiorities. Pretrained models consistently show higher median training accuracies, indicating more effective and stable learning across diverse architectures. Their higher test accuracies and lower variability also signal stronger generalization to new data. The uniformity in performance improvements across different architectures further accentuates the advantage of pretrained models. The notably lower median test loss of pretrained models underscores their enhanced efficiency in handling unseen data.