Exploring the potential of CNNs in detecting selection across temporally ancient and nearly neutral evolutionary scenarios in humans

Djinho Itshary
August 2024
Supervisor: Matteo Fumagalli

Contents

Abstract

The detection of natural selection within genomic data is crucial for understanding evolutionary processes, with significant implications for conservation biology, biomedicine, and agriculture. Traditional methods, such as likelihood-based approaches and Approximate Bayesian Computation (ABC), have become inadequate due to the computational intractability of calculating likelihood functions, reliance on summary statistics, and the curse of dimensionality. These challenges are exacerbated by the increasing size and complexity of genomic datasets driven by advancements in next-generation DNA/RNA sequencing technologies, making traditional methods insufficient for positive selection inference.

This study investigates the use of Convolutional Neural Networks (CNNs) as a model-agnostic method for detecting positive selection. CNNs, which are free from dimensionality issues, offer a promising solution to the challenges faced by the expanding size and complexity of genomic datasets. The study assesses whether CNNs can identify signatures of selection in temporally ancient and nearly neutral evolutionary scenarios, which have seldom been explored in the literature.

The findings of this study demonstrate that CNNs indeed possess the ability to detect temporally ancient and nearly neutral evolutionary scenarios, which previous methods relying on summary statistics have failed to achieve. After Bayesian optimization, the models were able to detect selection across all scenarios except one—recent and weak selection—where accuracy was limited to 50.40%. In contrast, for other equally challenging scenarios, the optimized model achieved accuracies ranging from 69.10% to 88.88%, with the recent strong scenario not requiring optimization due to an almost optimal baseline accuracy of 99.70%.

Additionally, this study introduces a novel application of CNNs by using trained models to test if autophagy-related disease genes, are under positive selection, aiming to apply the classifier to real genomic regions of clinical relevance. This has important implications for personalized medicine. Future work should focus on several key areas: expanding the dataset, particularly by increasing the number of batches used in training, conducting further experimentation with Bayesian optimization by running more trials with a more precisely defined hyperparameter space to improve model performance; and exploring transfer learning as a potentially faster route to achieving decent accuracy by leveraging pre-trained models. Also, implementing permutation-invariant models would improve the quality of the research in the future. Which would truly underscore the brilliance of CNNs as model-agnostic tool in detecting selection across complex evolutionary scenarios. All code for the project can be found at https://github.com/Djinho/EvoNet-CNN-Insight

1.0 Introduction

1.1 Background

The ability to detect natural selection within genomic sequences is an imperative task that significantly enriches our comprehension of biological and evolutionary processes. This critical capability holds profound implications for diverse scientific disciplines, including biomedicine and conservation biology. For instance, elucidating the selective pressures exerted on genomes is crucial for enabling us to predict how species may adapt and respond to climate change [1] providing critical insights for conservation strategies. Furthermore, it can elucidate the functional variants that underlie phenotypes associated with various diseases [2], thereby enhancing our understanding of the genetic contributions to disease pathogenesis. Moreover, being able to observe the selective pressures exerted on genomic regions helps us understand patterns of natural selection, which in turn makes it easier to elucidate the genomic foundations of evolutionary change in both wild and cultivated crop species [3]. Beyond that, this knowledge can inform breeding strategies to improve crop yield and resilience. In the context of human genetics, identifying instances of natural selection offers significant insights. For example, genes responsible for Mendelian diseases within the human genome are largely influenced by negative selection, while genes linked to complex traits may experience either negative selection or positive selection [4] and by identifying genomic regions currently under selection, researchers can potentially pinpoint genetic elements linked to illnesses or key traits [5], thereby offering significant implications for genetic research and personalized medicine. Moreover, as positive selection is the core instigator of the genetic adaptation in various species and in humans [6] understanding these processes can illuminate the mechanisms driving evolutionary change and adaptation.

1.2 Synopsis Of Summary Statistics And Methodological Frameworks For Selection Detection

Having established the significance of detecting natural selection within the genome, it is imperative to critically evaluate and give a brief overview of the contemporary methods available for this purpose, considering their historical usage within the literature and pitfalls. Contemporary techniques to elucidate signatures of selection predominantly rely on aggregate metrics, which condense information regarding the genomic variation of the population into concise summary statistics. Whereby the neutral distribution can be determined through empirical or analytical methods [7], albeit informative and computationally inexpensive [8] summary statistics lower the dimensionality of the genomic data [9] leading to a pertinent loss of information [11][10]. Which has been shown to reduce the statistical power [12]; being outcompete by other methods of selection inference, for example deep learning methods [13]. Moreover, the preponderance adaptive processes are driven by weak selection on pre-existing genetic variation [14]. As a result, many selective events produce signatures of selection that are subtle, intricate, and challenging to elucidate when relying exclusively on a limited range of summary statistics. Frameworks to detect selection can be categorised into two broad approaches: Likelihood methods and likelihood-free methods:

Likelihood Methods: These compute the likelihood function directly to infer selection. They rely on the explicit calculation of how likely the observed data is under a specific model [15][16] (see Figure 1). However, the computation of the likelihood function is computationally expensive [17]

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Figure 1: *Bayes' Theorem.*

Likelihood-Free Methods: These bypass the computation of the likelihood function, often due to its computational complexity, relevant examples are outlined below such as:

Approximate Bayesian Computation (ABC): ABC involves several key steps. Initially, it formulates various models and samples parameters from predetermined distributions (priors). For each model, data is simulated using these sampled parameters. Summary statistics are then calculated from both the simulated data and the empirical data. The final step involves comparing these summary statistics and retaining only those simulations that closely resemble the empirical data. [18][19] (see figure 2). ABC offers significant advantages over traditional likelihood-based approaches by circumventing the substantial computational burden of computing the likelihood function [20]. Instead of explicitly computing the posterior distribution of the estimated parameters, ABC approximates the likelihood function, achieving a model fit by generating simulations representing summary statistics that correspond to observed values [21]. However, a notable drawback of ABC is not only its reliance on summary statistics but also its computational intensity, as it requires numerous simulations. This can lead to what is known as the "curse of dimensionality" wherein the accumulation of sufficient information through an increasing number of summary statistics subsequently escalates the approximation error [21], meaning if the dataset is too large this can be a hinderance rather than a benefit.

Machine Learning Methods: While some machine learning methods still rely on summary statistics, they approach the detection of selection by training algorithms on simulated data to classify or predict patterns in real data. This reframes the task of selection detection as a classification problem rather than a traditional task of statistical inference (See Figure 2). However, not all machine learning approaches depend on summary statistics. For instance, deep learning techniques such as CNNs can directly analyse raw genetic data, bypassing the need for summary statistics altogether [13]. These methods are classified under the category of approaches that rely on simulations without depending on likelihood estimations (Cranmer et al., 2020) [20].

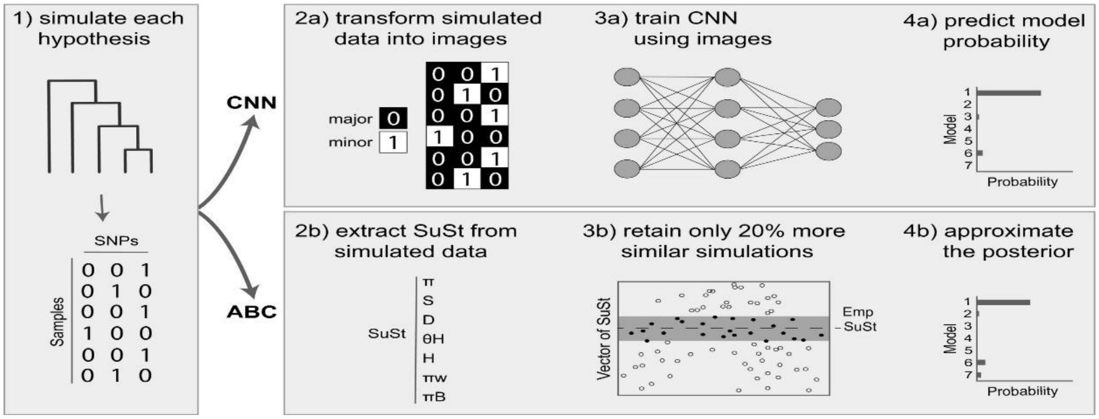

1) simulate each hypothesis
SNPs
Samples
0 0 1
0 1 0
0 0 1
1 0 0
0 0 1
0 1 0
CNN
ABC
2a) transform simulated data into images
major 0
minor 1
3a) train CNN using images
4a) predict model probability
Model
1 2 3 4 5 6 7
Probability
2b) extract SuSt from simulated data
SuSt
π
S
D
θH
H
πW
πB
3b) retain only 20% more similar simulations
Vector of SuSt
Emp SuSt
4b) approximate the posterior
Model
1 2 3 4 5 6 7
Probability

Figure 2: *Comparison ABC and ML Approaches.* The figure illustrates the process of ABC, which involves sampling parameters, simulating data, calculating summary statistics, and retaining simulations that closely match empirical data. It also shows how machine learning methods, particularly convolutional neural networks (CNNs), can circumvent using summary statistics by directly analysing raw data. Adapted from Perez et al. (2020) [18].

2.0 Literature Review

2.1 Evaluation Of Single Summary Statistics

Given the centrality of summary statistics in many selection inference methods, it is essential to outline and critically evaluate them, particularly in terms of their underlying biological assumptions. In this context, I will provide an outline of the early summary statistic methods that have been developed over the past several decades, focusing on their evolution and application in detecting natural selection, while also highlighting their limitations and assumptions.

Early summary statistic-based methods were formulated to identify strong selective sweeps, by examining their impact on the site frequency spectrum using summary statistics such as “Tajima’s D” [22] and FU Y.X “H” statistic [23]. These methods were predicated on the prevalence of derived alleles at both low and high frequencies, to detect high frequency variants as a signature of directional selection within the genome [22][24]. However, these SFS-based methods were observed to be highly susceptible to confounding factors. For instance, “Tajima's D,” as used in the study by F Tajima [22], is sensitive to deviations across the entire frequency spectrum. Additionally, the “H” statistic is only effective in detecting a significant surplus of high-frequency variants [25]. Moreover, this class of summary statistics is susceptible to confounding, particularly when dealing with complex demographic models, such as those involving population expansion and bottlenecks [26]. This presents a major difficulty in selection inference as it confines the application of demographic models to equilibrium conditions, despite a substantial portion of demographic histories involving non-equilibrium scenarios, such as expansions and bottlenecks [27]. There have been a preponderance other studies within the corpus of the literature that have used SFS as a metric to infer selection [28] [29][30] [31] and have all had the common issues and pitfalls as mentioned above.

Other studies have leveraged the presence of amplified linkage disequilibrium at the specific genomic site of a genetic sweep, identified through correlations in allele frequencies [32][33]. While linkage disequilibrium measures provide valuable insights, they have notable limitations. LD is extremely sensitive to the sample's genealogical structure, which can lead to a lot of variation even when conditions are neutral. This can make it less powerful to reject the neutral model unless the differences are noticeably big [32]. Moreover, in regions with high recombination rates, the rapid decay of LD further limits its utility. Furthermore, differentiating the signatures of selection from the signatures of recombination presents a significant challenge, as both processes can generate similar allele frequency patterns. This overlap complicates interpretation and necessitates the use of additional models to accurately account for these intertwined evolutionary dynamics [32]. Moreover, methods that incorporate LD have been shown to perform only slightly better than SFS methods, such as Tajima's D and others previously mentioned. This marginal improvement is likely due to the correlation between LD and SFS [33]. One could argue that the implementation of these LD summary statistic methods for detecting selection may not justify the added bias and variance, compared to previously employed SFS-based summary statistic methods.

Lastly, these methods rely on several critical assumptions, such as a stable population size, the recent emergence of the advantageous mutation, a non-zero recombination rate, genetic isolation, and no gene conversion near the fixed allele [34][35]. However, only a limited number of statistical models fully satisfy all five of these assumptions, underscoring the necessity for more versatile and practical methodologies to detect selection signatures [36].

Given these limitations, traditional single summary statistic-based methods, whether leveraging SFS or LD, are constrained by their underlying assumptions and are vulnerable to confounding factors. Consequently, there is an urgent need for more robust and adaptable approaches capable of addressing the increasing complexity and scale of genomic datasets resulting from advances in Next-Generation DNA/RNA sequencing technologies [37][38]. This need has driven the integration of machine learning techniques, which offer a novel framework by reframing selection detection as a classification problem rather than a traditional statistical inference task, which will be discussed below.

2.2 General ML Methods For Detecting Selection

This section provides a critical review of the integration of machine learning methodologies within evolutionary genomics, with a particular focus on detecting natural selection. The advent of machine learning represents a significant shift from traditional approaches, such as single and ensemble summary statistic methods and Approximate Bayesian Computation (ABC), which are often limited by information loss, the curse of dimensionality, and the challenges posed by intercorrelated summary statistics [11][21][39][40]. We begin by surveying general machine learning techniques employed in this domain, highlighting their contributions and limitations. Subsequently, we will concentrate on deep learning, particularly convolutional neural networks (CNNs), a sophisticated and increasingly prominent faction of supervised ML algorithms rooted in artificial neural networks [41]. Given the importance of establishing a comprehensive framework, this section also traces the gradual progression of machine learning applications in selection inference, culminating in the rationale for incorporating CNNs as the core model in this project.

The earliest application of machine learning for selection inference was conducted by Pavlidis et al. [42], who used Support Vector Machines (SVM) to infer positive selection from SNP data, particularly in non-equilibrium populations. Their approach was combined with the ω statistic, introduced by Kim and Nielsen [33], which measures the spatial distribution of LD around a selection scan using the composite likelihood ratio (CLR), and with SweepFinder [43]. Pavlidis et al. [42] made several modifications to these methods, including the implementation of a variable ω statistic as he noticed that a large ωMAX can attenuate signals of positive selection, and a small one may miss SNPs. Additionally, he introduced a modified SweepFinder that includes a fraction of monomorphic sites to improve the accuracy of identifying the presence of selection by preserving the signature of reduced genetic diversity, thereby enhancing mutation detection realism [42]. The modified ω-statistics and SweepFinder outputs were used as input features for the SVM classifier. The SVM outperformed both the modified ω-statistic and SweepFinder approaches, even under demographic models that are not in equilibrium. However, it was found that the complexity of implementing the ω statistic made the machine learning approach computationally intractable for large datasets [44].

Building on previous SVM-based approaches for detecting positive selection, Ronen et al. [45] incorporated the use of SVMs to analyze site frequency spectrum data for the detection of selective sweeps, leading to the development of "SFselect," a model rigorously trained on large, simulated datasets using the "mpop" forward-in-time simulator [46]. “SFselect” outperformed existing methods, demonstrating its effectiveness when applied to real genomic data from mosquitos and humans. The analysis identified novel loci under directional selection linked to hypoxia tolerance in mosquitos and pigmentation adaptation in humans. The study further identified the specific summary statistics that excel at detecting selection across different temporal scales, outperforming previous approaches, including the method previously described by Pavlidis et al. [42], as mentioned earlier in this paper. However, a notable limitation of the study is that the effectiveness of the SVM depends on the specific parameter space used for training, which may limit the model’s applicability to untested scenarios. Despite this, SFselect demonstrated strong effectiveness in detecting positive selection across various selection coefficients and demographic contexts.

Boosting, a technique that combines multiple weak models—each slightly better than random guessing—into a highly accurate predictive model [47], was employed for selection detection, as demonstrated by Lin et al. [49], who developed the R package EvolBOOSTING [48]. The researchers utilized simulated data to optimize an ensemble of summary statistics (e.g., Tajima's D, iHH) for distinguishing between selection, neutrality, and demographic events such as bottlenecks [49]. This method outperformed the previously mentioned approaches, including those developed by Pavlidis et al. [42] and Ronen et al. [45], achieving higher power in detecting weaker selection. However, a notable limitation is the potential for false positives for demographic events, such as bottlenecks, as they resemble selective sweeps. This issue was mitigated by employing a two-step classification process inspired by Thornton and Jensen 2007 [50]. However, a notable limitation of the method is its dependency on location specificity. Specifically, when the classifier is trained on a particular site, its accuracy diminishes significantly if the testing sample originates from a different location

Following this, hierarchical boosting was implemented to infer selection, as demonstrated by Pybus et al. [52], who utilized the *mboost* R package [51]. Their model categorizes genomic regions by analyzing summary statistics from various tests for positive selection. The model was developed using coalescent simulations based on human demographic data and validated with data from three populations in the 1000 Genomes Project. The study effectively identified selective sweeps, categorizing them based on their completeness and age and not just as a binary outcome (e.g. selection vs neutral selection) compared to the previously mentioned methods, the hierarchical boosting approach outperformed those described by Ronen et al. [45] and Lin et al. [49] in detecting both recent and ancient selection. A key finding was the reduced number of sweeps observed in African populations, which the study postulated was like due to demographic factors and African populations more likely undergoing standing variation (soft sweep), rather than de novo mutation (hard sweeps).

Building on this, Schrider et al. [53] developed the Soft/Hard Inference through Classification (S/HIC) method, a machine learning algorithm that uses an Extra-Trees classifier to distinguish between hard and soft selective sweeps. By incorporating an ensemble of summary statistics across genomic windows, S/HIC demonstrated high accuracy across diverse demographic scenarios, even when models were mis specified. Notably, it effectively detected both known and novel sweeps in the 1000 Genomes CEU dataset. Moreover, S/HIC achieved the highest ROC curve scores for detecting both soft and hard sweeps, surpassing the performance of previously established methods such as SFselect [45], EvolBOOSTING [49], SweepFinder [42], and Tajima's D [22].

2.3 The Application of DL and CNNs in Detecting Natural Selection

Having examined preliminary machine learning approaches such as Support Vector Machines (SVMs), boosting, and ensemble methods—each of which has significantly advanced the task of positive selection inference —the next frontier lies in deep learning. This project specifically focuses on Convolutional Neural Networks (CNNs), which present new opportunities for innovation in evolutionary genomics [54] due to its raw data extraction capabilities. Therefore, it is essential to critically review and evaluate the application of CNNs for selection inference, identifying any gaps in the literature, in which my project seeks to fill. By assessing the effectiveness of CNNs in detecting positive selection, this work aims to build upon existing methodologies while highlighting the unique advantages and challenges associated with CNNs in this context [55].

The earliest applications of deep learning in evolutionary genomics to infer positive selection marked a significant shift from traditional methods. Sheehan and Song (2016) [56] pioneered this approach with the introduction of “evoNet,” a deep feedforward neural network trained on simulated data. Although "evoNet" outperformed traditional ABC methods in detecting selection and inferring demography, it did not employ CNNs but rather relied on a feedforward neural network. This approach utilized summary statistics as input, which, while effective, did not fully exploit the raw data extraction capabilities that deep learning employs has.

Building on his previous work with S/HIC, which initially employed a different ML algorithm, Schrider et al. [57] developed “diploS/HIC”, marking a significant innovation in the application of CNNs in the task of positive selection inference. While other approaches in the field have utilized CNNs to detect positive selection using raw summary statistics, Schrider's method uniquely transforms these summary statistics into image data. Specifically, 'diploS/HIC' encodes summary statistics in a matrix format, where the rows denote various statistics and the columns represent sub-windows, forming an image representation of a genomic locus. This novel approach leveraged the strengths of CNNs to capture complex patterns in the data, advancing the detection of positive selection beyond traditional methods. Moreover, diploS/HIC outperformed Schrider's earlier methods that utilized an Extra-Trees classifier for both phased and unphased genomic data. However, similar to prior approaches, “diploS/HIC” still depended on summary statistics rather than raw genetic sequences, despite converting these statistics into image data.

To maximize the data extraction potential of CNNs, a more advanced approach was explored by replacing summary statistics with raw sequence alignments as input. A landmark study by Torada et al. [13] exemplifies this innovation. Using simulated data, they simulated haplotypic images representing neutral selection and selective sweeps. These images underwent preprocessing, such as removing rare variants and applying data augmentation, to generate a binary matrix of haplotypes, which the CNN used as input for the first layer. Torada et al. [13], even found that sorting rows by frequency improved performance, particularly with small sample sizes. It is important to note that while data preprocessing and augmentation enhances image clarity, this raises a critical concern: are we training models to detect only the patterns we expect? [58] which could raise doubts about whether this approach truly utilizing the model-agnostic nature of CNNs.

Further advancements in using CNNs to detect signatures of positive selection are demonstrated in recent studies. For instance, Ulas et al. [59] implemented CNNs to identify recent balancing and positive selection, showing that CNNs significantly outperformed their predecessor, the fully connected neural network (FCNN), in making these genomic inferences. Additionally, CNNs have consistently shown superior performance over summary statistic-based methods in detecting positive selection, even with relatively simple and shallow models. This is exemplified by the work of Cecil et al. [60], who extended the framework of Torada et al. [13] by developing a 'mini-CNN’. Whereby, despite its simplicity, the mini-CNN maintained high efficacy in identifying recent and strong selective sweeps [60]. Moreover, it’s reduced complexity also provided greater transparency into the model's decision-making process, addressing concerns of the "black box” nature of CNNs [61] as it is an inherently arduous task understanding the inner workings and decision rules of a CNN. Notably, Cecil et al. [60] discovered that the mini-CNN primarily focusses on variations between rows, with the top rows of the image contributing significantly to the model’s predictions. In addition, Cecil et al. [60] also tested a permutation-invariant CNN architecture known as 'DeepSet,' which represents a step closer to achieving a truly model-agnostic approach. Which addresses some of the challenges previously identified by Flagel et al. [58] on whether we are training models to perceive what we want them to perceive. However, a notable drawback of the study by Cecil et al. [60] is the overgeneralization of the CNN performance based on a single case, which overlooks evidence that some CNNs outperform traditional methods. This led to a questionable claim that Garud's H1 [62] outperforms CNNs, and a misleading assertion that CNNs merely mimic summary statistics through row sorting.

Lastly, Fadja et al. [63] expanded on the work of Torada et al. [13] , as they tested a CNN on different dataset sizes, with a dataset of 50,000 training images, 10,000 validation images , and 10,000 testing images achieving the best accuracy. To validate their model, they applied it to real data from the SLC24A5 gene, linked to skin pigmentation in Europeans, achieving 88% accuracy. This application to real genomic data was a novel implementation compared to similar tasks in the past that did not incorporate real data testing.

3.0 Fundamental Basics of CNNs

Building upon the previous discussion of machine learning techniques for inferring selection, CNNs were introduced as a key tool in this domain as the next frontier in this this task. It is important to delve deeper into the architecture of CNNs. Given that CNNs will be employed in this study, it is pertinent to explain their fundamental components and operational mechanisms. CNNs are composed of several key layers, including the convolutional layer, rectified linear unit (ReLU), pooling layer—specifically Max-pooling in this study—and the fully connected layer [130]. Understanding these components is crucial for the effective implementation of CNNs in the study of selection inference, as they collectively enable the network to learn and generalize from data in a hierarchical and efficient manner. Additionally, the model-agnostic nature of CNNs allows us to learn new insights from nature in a model agnostic way [55] through its capacity for raw feature extraction, making it a powerful tool for uncovering patterns and relationships within complex datasets.

3.1 Convolutional layer

This layer is responsible for the first stage of feature extraction, which generally involves applying both linear and non-linear operations [130]. The process involves employing a kernel to slide across the input tensor, where the kernel performs element-wise multiplication and summation, producing a single value for each position on the corresponding feature map (Figure 16). This operation is known as a convolution and is repeated several times. As a result, a feature map is created, displaying all the extracted features and patterns from the input tensor. The feature map will then be pooled, which will be explained below.

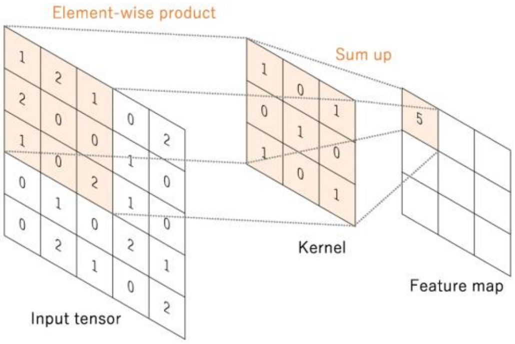


Figure 16: Convolution operation in a neural network where the kernel performs element-wise multiplication with the input tensor, and the results are summed to form the feature map. Image from Yamashita et al (2018) [130].

3.2 Rectified Linear unit

ReLU is one of the most utilized activation functions in neural networks [131]. ReLU is a type of ramp function that outputs the input value directly if it is a positive, and zero if it is negative. This straightforward approach offers significant computational efficiency compared to other activation functions like sigmoid or tanh, and it also mitigates the vanishing gradient problem [132].

3.3 Pooling layer

Pooling reduces the dimensionality of the input tensor. This reduction not only decreases the number of parameters but also lessens the computational complexity of the model [133]. By down sampling the feature maps, pooling enhances the model's robustness to small shifts and distortions [130]. There are various types of pooling operations, but for conciseness, I will focus on max pooling, which is utilized in this study. Max pooling involves selecting the highest value from each sub-region of the input tensor, as depicted in Figure 17. This method keeps only the most prominent features by selecting the highest value from each subregion, which helps maintain key information while lowering the dimensionality [130]. Additionally, including more pooling layers can help detect finer details in the data [133].

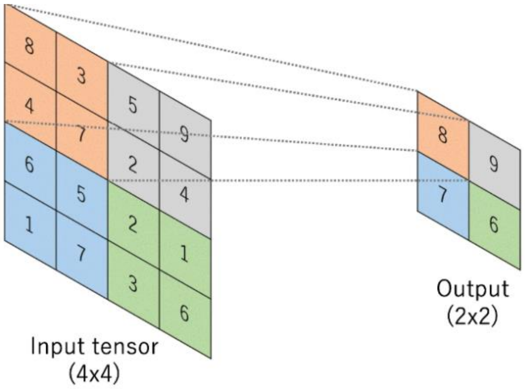


Figure 17: outlining the basics of the max pooling operation, where the maximum value from each 2x2 region of the input tensor is selected to form the output tensor. Image from Yamashita et al (2018) [130].

3.4 Fully connect layer FCNN

This segment of the CNN flattens the feature maps from the antecedent layers into a 1D array, enabling each element to connect to every neuron in the fully connected layer with distinct weights [130]. In the FCNN, each neuron is linked to all activations from the previous layer. The FCNN then integrates the features learned in earlier layers to make a final prediction, often providing a probability value for each class in a classification task [130]. The last fully connected layer typically contains as many nodes as there are classes and applies an activation function such as sigmoid for binary classification tasks or softmax for multi-class classification tasks.

4.0 Motivations For The Study

The implementation of deep learning in other realms of genomics has shown great promise, such as elucidating high frequency recombination sites [64] uncovering adaptive gene transfer between populations [65] and identifying differences between soft sweeps and balancing selection [59]. These methods represent a significant advancement by fully leveraging raw data and eschewing reliance on summary statistics, thereby yielding more precise and comprehensive insights into evolutionary processes. However, the application of CNNS in evolutionary genomics for selection inference remains relatively unexplored [55]. This study aims to fill this gap in the literature by leveraging CNNs to investigate selection processes that span extensive temporal scales and nearly neutral evolutionary scenarios, which contemporary research seldom addresses. Most existing studies exhibit temporal myopia, focusing on recent, strong, and soft sweeps, and rarely extending beyond selection events that occurred more than 20,000 years ago [13][60][63]. Some studies have combined CNNs with summary statistics with the efforts of inferring ancient selection events (those occurring more than 1000 generations ago). While these approaches are informative, they still rely heavily on summary statistics [66]. Which don’t fully utilize the raw data extraction capabilities that CNNs possess.

My research seeks to fill critical gaps in our understanding of how key evolutionary periods have influenced the genetic landscape of modern humans by evaluating the capacity of CNNs to detect selection in temporally ancient and nearly neutral evolutionary scenarios. Specifically, the study will assess whether CNNs can infer selection across temporal distances of 100,000, 50,000, and 10,000 Kya, which are periods marked by significant events that have profoundly impacted human evolution, which will be discussed in further detail. For instance, around 100,000 years ago, the initial migration of Homo sapiens out of Africa was driven by climatic changes that created habitable corridors [67]. This was followed by a larger wave of migration approximately 50,000 years ago, which further spread human populations across Eurasia [68] and introduced new genetic diversity through interactions with other hominids, such as Neanderthals and Denisovans [69][70]. By 10,000 years ago, during the Holocene, the onset of agriculture and animal husbandry marked a major shift from hunter-gatherer societies to settled agricultural communities, leading to significant genetic adaptations [71] due to change in diet and cultural practices e.g., diary consumption.

My research aims to clarify how natural selection during ancient migrations out of Africa has shaped the genetic diversity of modern humans. Specifically, it focuses on how founder populations, encountering new climatic pressures, dietary shifts, and cultural changes during the late Pleistocene and early Holocene have impacted the human genome. The use of CNNs in this context represents an innovative approach to understanding evolutionary dynamics over extensive temporal scales, addressing a crucial gap in contemporary research. This study introduces novel optimization strategies to identify motifs in specific architectures suited for these evolutionary scenarios. Utilizing the parametric software ImaGene, we aim to detect the presence of selection by generating datasets based on the Northern European demographic model, across varying selection strengths, within the precise time scales outlined above. The ImaGene software will provide the baseline assessment of the CNN architecture as the software utilises a low-complexity CNN at baseline, which will then be optimized using Bayesian methods to enhance detection capabilities. This approach explores the potential for more refined detection of selection signatures, contributing to a broader interpretation of human evolutionary history and will allow us to assess whether low complexity CNN architectures can infer ancient and nearly neutral selective events and whether complex architectures can do the same.

Moreover, there has been limited guidance on which types of CNN architectures are most effective for detecting specific types of positive selection, particularly across different temporal epochs and selection strengths in evolutionary genomics. Previous studies have experimented with architectural variations to develop more efficient CNNs, but these efforts have been largely confined to detecting recent and strong selective sweeps [60]. Other research has focused on dataset size [63] and data augmentation [13], yet no study has explicitly outlined the architectural motifs required for detecting different forms of selection. For example, it has not been clearly established that ancient selection might benefit from deeper and narrower architectures, such as six layers with filters ranging from 64 to 128.

Furthermore, there is a conspicuous lack of research investigating whether genes linked to diseases characterized by autophagy dysregulation are subject to positive selection. This disparity is particularly notable given that other genes associated with disease have been tested for selective pressure. A prominent example is the positive selection of specific genes related to hemoglobinopathies, such as the HbS allele associated to sickle cell anemia, which confers substantial immunity against malaria. The HbS allele, known as the 'sickle cell trait,' is a well-documented case of positive selection driven by exposure to infectious disease, offering significant survival benefits against severe malaria in the heterozygous condition. However, in the homozygous state, it increases the risk of sickle cell anemia [72][73][74]. Additionally, BRCA1 has also been elucidated to be under positive selection [75] [76] and this gene has been linked to increase risk of breast cancer and ovarian cancers [77]. Moreover, genomic scans have identified the gene UGT2B4, which is linked with an elevated risk of breast cancer in Nigerians, and as being under both balancing and positive selection [78].

So, we have extensive records of identifying disease-associated genes under positive selection, which raises the question: why not investigate autophagy dysregulation genes in a similar manner? These genes are linked to several malignancies and chronic conditions. For instance, ankylosing spondylitis, which affects approximately 31.9 people per 10,000 in North America, is associated with autophagy-related pathways [79] [80]. Similarly, hypertension, now affecting over 1.13 billion adults globally, also shows links to autophagy dysregulation [81][80]. Breast cancer, which led to 2.3 million diagnoses and 666,000 deaths in 2022, involves genes associated with autophagy dysregulation [82][80]. Additionally, Selective IgA deficiency, affecting 1 in 600 people, is another condition in which autophagy dysregulation genes play a crucial role. The common denominator between all these malignancies is that they have pathogenic variants located on the ULK1-Ccomplex [80] (see figure 3). Given the widespread prevalence and severe impact of these malignancies, investigating whether the associated genes are under positive selection could provide valuable insights into the evolutionary processes and selective pressures that have shaped them. Understanding the selective history of these genes may also contribute significantly to the advancement of personalized medicine, particularly in tailoring prevention and treatment strategies for populations at higher risk [5]. Notably, this research is pioneering in its application of CNNs to detect whether disease-related genes are under positive selective pressure—a task that has not been previously undertaken. Unlike traditional studies that often rely on specific hypotheses or known selection markers, this approach tests for selection without any prior assumptions, allowing for an unbiased exploration of evolutionary patterns. This research has the potential to uncover new targets for therapeutic intervention, enhance our understanding of disease etiology, and improve our ability to predict disease risk based on genetic profiles (specifically Northern Europeans, in the context of this study).

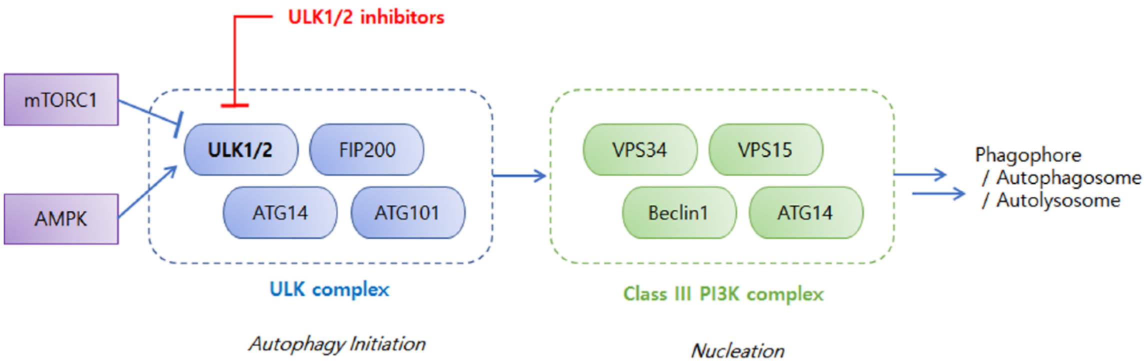


Figure 3: Overview of the autophagy initiation pathway, highlighting the ULK1 complex (ULK1/2, FIP200, ATG14, ATG101) regulated by mTORC1 and AMPK. ULK1/2 inhibitors (red) block this complex, which in turn activates the Class III PI3K complex (VPS34, VPS15, Beclin1, ATG14) essential for nucleation and the formation of autophagosomes. Adapted from Karmacharya and Jung (2023) [123].

4.1 Research Objectives And Aims

This research aims to develop and evaluate CNNs for their ability to accurately detect temporally ancient and nearly neutral evolutionary scenarios, striving to achieve industry-standard accuracy. To enhance performance, Bayesian optimization will be employed to refine the CNNs in the hope of increasing validation accuracy and identifying architectural motifs. Trained models will be cross validated using known genetic variants under positive selection to ensure robustness. Additionally, an exploratory analysis will be conducted to investigate whether genetic variants related to autophagy dysregulation within the ULK1 complex are subject to positive selective pressure.

5.0 Methods

This section details the study's methodology, including experimental design, data preprocessing, and optimization techniques. Each choice is rigorously justified, demonstrating alignment with the study's objectives and enhancing the reliability of the results.

5.1 General Overview of the methods

This study employs a dual-phase experimental design to rigorously evaluate and optimize CNNs for detecting positive selection. The process begins with assessing a baseline low complexity CNN architecture which is offered by ImaGene, followed by hyperparameter optimization through Bayesian Optimization to enhance performance. The evolutionary scenarios encompass nine distinct datasets representing a matrix of evolutionary scenarios across diverse selection start times and selection coefficients (See Table 1).

Table 1: This table summarizes evolutionary scenarios based on different selection intensities (S = 100, 200, 300) and timing (Recent 10 Kya, Intermediate 50 Kya, Ancient 100 Kya).

	Evolutionary Scenarios		
	Recent 10kya	Intermediate 50kya	Ancient 100kya
S = 100 (Weak)	Recent, Weak	Intermediate, Weak	Ancient, Weak
S = 200 (Moderate)	Recent, Moderate	Intermediate, Moderate	Ancient, Moderate
S = 300 (Strong)	Recent, Strong	Intermediate, Strong	Ancient, Strong

Note. S" denotes the selection coefficient: Weak (S = 100), Moderate (S = 200), and Strong (S = 300). Timeframes indicate when selection events occurred: Recent (10 Kya), Intermediate (50 Kya), and Ancient (100 Kya).

Following the optimization process, CNN models that achieve a validation accuracy of 70% or higher will be considered fit for purpose. This threshold aligns with industry standards for novel binary classification tasks [84]. Achieving this accuracy indicates that the model has learned important motifs within the dataset and is significantly better than random chance at classifying an image [85]. Once the models meet the initial criteria, they will be further tested using known candidate genes under positive selection. This evaluation will assess the model's ability to predict genes under positive selection and to infer the timing and strength of these selection events. The predicted class probabilities from the models will be compared with established literature to evaluate their accuracy and reliability. For example, if a model trained on ancient-weak selection scenarios yields a high probability for a gene region known to have undergone similar selection, it would validate the model's effectiveness. Additionally, the models will be tested on disease genes known to be linked to autophagy dysregulation, particularly involving the ULK1 complex. This comparative analysis not only ensures that the models align with documented evolutionary history and selection pressures but also holds the potential to uncover new insights about genes that have never been assessed for directional selective pressure.

5.1 Experimental Framework

This project utilizes ImaGene, a user-friendly software designed for deep learning applications in bioinformatics. ImaGene allows users to select parameters and create evolutionary scenarios, using population genomic data for CNN-based classification tasks [13]. The software processes genomic data through multiple stages to ensure data quality and consistency, which aids in generating visual representations suitable for CNN analysis.

ImaGene Machine learning pipeline in short comprises the following steps:

1. **Simulations**: Perform simulations conditional on demographic models and selection events where selection start time and intensity can be varied, producing training and testing sets.
2. **Image Representation**: Converts genomic data into images by translating population structures and allele frequencies into a format understood by CNNs, which are haplotype matrixes.
3. **Prediction and Quantification**: Utilise Keras [86] and TensorFlow [87] for training and testing the CNN models. Post-training, these models are used to predict outcomes by generating probability distributions for the parameters of interest. In this experiment, the focus is on estimating the selection coefficient and determining the presence of selection within a given genomic region.

5.3 Dataset Generation

The datasets used for this analysis were created using msms [88], a coalescent simulation program that operates backwards in time. This method was coined and developed by Kingman (1982) [89]. The fundamental idea is that the simulations trace the lineage of people backwards in time until all samples meet at a shared ancestor. Kingman defined this process as the estimation of the probability of lineages merging or recognizing common ancestry within a population that mates randomly [89]; msms allows us to simulate genetic data under diverse demographic models, capturing details on genetic ancestry, recombination events, mutations, and population demography. The resulting output files generally contain ancestral recombination graphs, genealogies, and sequences, offering a thorough depiction of the simulated genomic regions [88]. The specific simulation parameters of the datasets represent a matrix of evolutionary scenarios (See Table 1).

Selection coefficients in msms quantify the intensity of positive selection acting on specific alleles. These coefficients are mathematically represented as $2N_{es}$, where N_e denotes the effective population size and s is the selection coefficient [90]. For this study, selection coefficients of 100, 200, and 300 were employed to represent weak, moderate, and strong positive selection pressures, respectively. These coefficients were applied to different genotypes: SAA for homozygous dominant (AA), SAa for heterozygous (Aa), and Saa for homozygous recessive (aa). Strong positive selection, indicated by a coefficient of 300, suggests a substantial fitness advantage for the allele in question. This advantage results in the rapid increase in the frequency of the advantageous allele within the population. As this allele becomes more prevalent, the genetic diversity within the surrounding chromosomal region diminishes, leading to a reduction in haplotype and nucleotide diversity [91]. This phenomenon is referred to as a 'selective sweep,' where the advantageous allele, along with linked neutral or even slightly deleterious variants, spreads through the population, leaving a signature of reduced genetic variation in the affected genomic region [92]. Figure 4 visually outlines what we are trying to detect or infer as a signature of positive selection by the CNN.

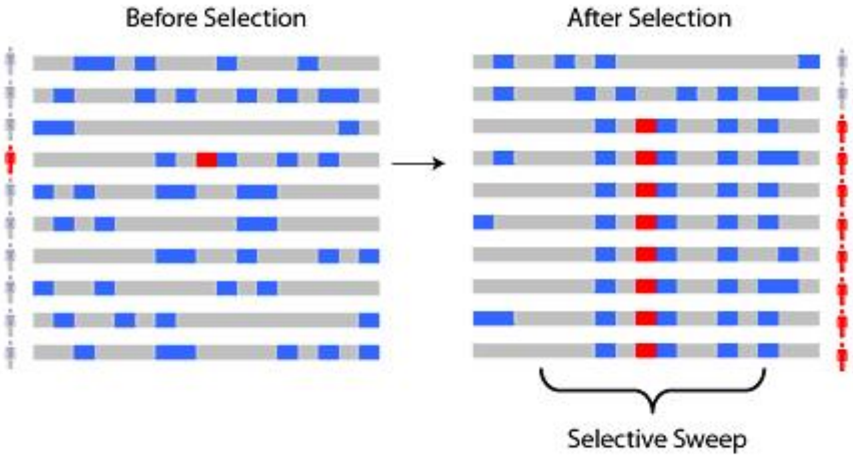


Figure 4: Illustration of a selective sweep before and after selection. Before selection (left), genetic variants (blue and red blocks) are evenly distributed across individuals. After selection (right), the advantageous red variant increases in frequency, leading to reduced genetic diversity around the selected locus. This process results in a selective sweep, as shown by the clustering of red blocks. Adapted from Schaffner and Sabeti (2008) [124].

Temporal measurements in coalescent simulations are expressed in units of 4Ne generations. For example, a time parameter of 0.1 corresponds to 100,000 years ago, given an effective population size of 10,000 and a generation time of 25 years. To convert years to generations, divide the number of years by the generation time. Thus, 100,000 years divided by 25 years per generation equals 4,000 generations, meaning a selection event 4,000 generations ago occurred 100,000 years ago. In the ImaGene parameter file, this is represented as 0.1, or as 4000/4000 using Bash/Linux-based software. Similarly, for 50,000 years ago, the representation is 2000/4000 (or 0.05), and for 10,000 years ago, it is 400/4000 (or 0.01).

The datasets consisted of 40,000 simulations per evolutionary scenario, with 2,000 images per class (selective sweep or neutral) in each batch, resulting in 4,000 images per batch. With 10 batches of data, this totaled 40,000 simulations per dataset. The number of simulations per dataset strikes a balance between computational efficiency and providing sufficient training data for model accuracy. While research has shown that as few as 10,000 simulations can yield accurate results for CNN based classification tasks [18]. More contemporary research has employed upwards of 300,000 simulations [65]. By selecting 40,000 simulations per dataset, this study ensures robust training while maintaining computational feasibility.

The demographic model employed as a constant in these simulations is derived from Marth et al. (2004) [27]. Which encapsulates a sequence of population size alterations over time. The model parameters are delineated as follows: DEMO='-eN 0.0875 1 -eN 0.075 0.2 -eN 0 2'. This parameter set articulates a demographic history characterized by temporal fluctuations in population size. Specifically, at 0.0875 units of $4N_e$ generations in the past, the population size is initialized to 1. Subsequently, at 0.075 units of $4N_e$ generations ago, the population size is reduced to 0.2. At time 0 (representing the present), the population size is increased by a factor of 2 (See figure 5). This model captures substantial historical demographic shifts in the Northern European population spanning back 87,500 years ago. Notably, this demographic model was maintained as a constant throughout all datasets.

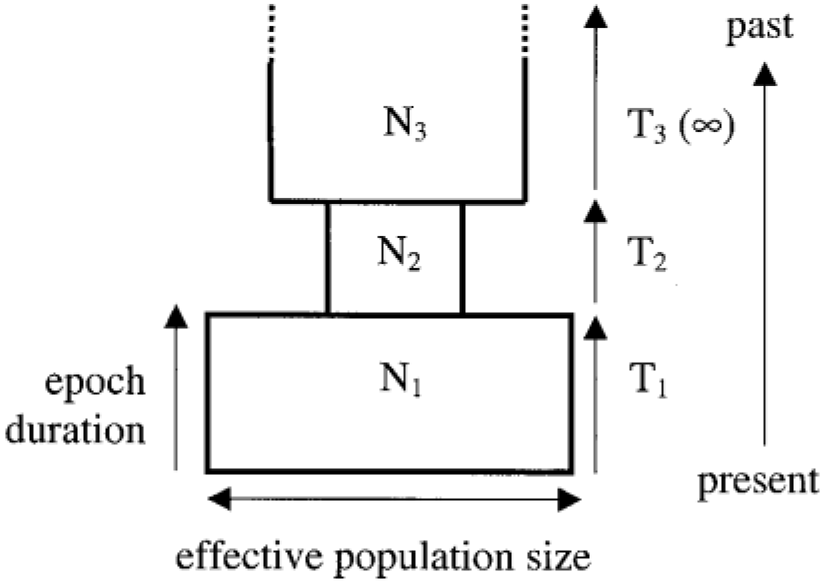


Figure 5 : This illustration consistently represents the demographic model used in our experiment. It depicts a three-epoch population history characterized by a bottleneck event with a piecewise constant structure. Initially, the ancestral effective population size (N_3) undergoes a sudden reduction to N_2. This intermediate size is maintained for T_2 generations before a stepwise increase to N_1, occurring T_1 generations before the present (Adapted from Marth et al. 2004 [27]).

The simulations were conducted with a constant locus length of 80,000 base pairs, using a mutation rate (θ) of 48 and a recombination rate (ρ) of 32, corresponding to 1.5×10^{-8} and 1×10^{-8} per base per generation, respectively, consistent with representational estimates for the human genome [93][94]. A sample size of 198 chromosomal copies was employed. Selective pressure was consistently applied to a variant centrally located within the specified region. These parameters were held constant across all simulations to ensure consistency and reliability, enabling the isolated assessment of the effects of varying only the selection start time and strength

5.3.1 Data Preprocessing

To enable the effective training and testing of the CNNs, each set of simulated haplotypes was converted into a haplotype matrix, which effectively serves as an image with fixed dimensions. In this format, each row represents a specific haplotype, and each column corresponds to a genomic location. We pre-processed the sampled haplotypes from each simulation. This pre-processing included converting the haplotypes into binary images through major/minor allele polarization, eliminating loci with allele frequencies that were too low, ensuring uniform width across all images and grouping the rows into common haplotype segments and arranging these segments according to their frequency. Prior to the data processing, the haplotypes look like those illustrated in Figure 6, raw genetic data where rows represent samples and columns represent loci. The black and white pixels indicate different alleles, forming a chaotic, hard-to-interpret pattern

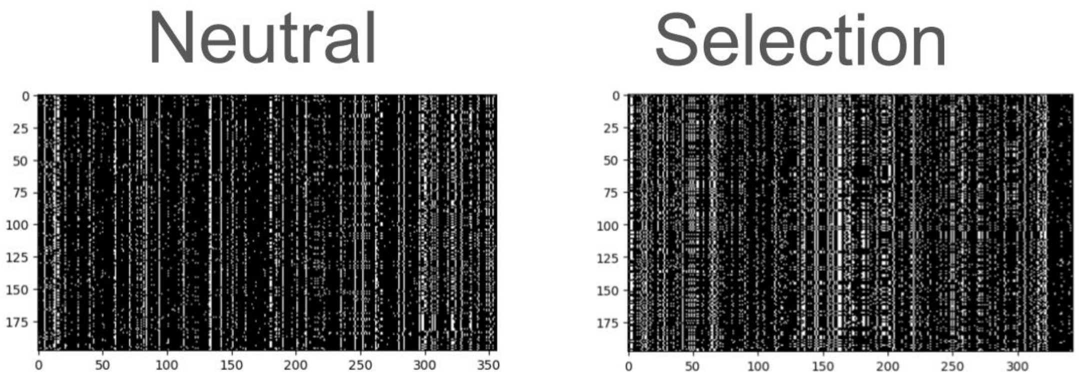


Figure 6: Raw Haplotype Data under Neutral Evolution and Selection: The figure shows raw haplotypes from genetic simulations, with "Neutral" (left) and "Selection" (right) conditions. Rows represent samples, and columns represent genetic loci, with black and white pixels indicating different alleles. Patterns are more chaotic under neutrality and more structured under selection.

5.3.2 Frequency Filtering

Genetic variants were meticulously filtered to retain those with a minor allele frequency (MAF) above 1% (0.01). This threshold was chosen to eliminate noise, remove rare variants and focus the model on relevant genetic variants more likely under selection pressure [95][13] and will speed up the training time.

5.3.3 Sorting and Resizing

To aid in pattern recognition, binary matrices representing genetic variants were sorted by allele frequency. Each matrix was then resized to a standardized dimension of 198×192 with a single RGB color channel of 1 (to represent greyscale images) using the skimage resizing algorithm in Python. This process ensured consistent image dimensions for all images. This normalization of input shapes is crucial for ensuring compatibility and uniformity in CNN processing [96]. Moreover, this step was done to match the dimensions of the real data we will be testing. Following this, the rows within each matrix were further sorted by the frequency of their occurrence, grouping similar haplotypes together. This additional step enhances the clarity of genetic patterns, making them more distinguishable and has been proven to improve the model's prediction accuracy, particularly in cases where selection coefficients are low [13][58].

5.3.4 Image Flipping

Following sorting and resizing, the haplotype images underwent data augmentation through pixel inversion (flipping). This process not only increased the diversity of the training set [97] but also enhanced the visual distinction of genetic patterns, making the images more suitable to be a CNN input. After completing the aforementioned data preprocessing steps, including flipping, the input data resemble the images displayed in Figure 7.

5.3.5 Subset Selection and Target Definition

Random subsets of the data were selected to ensure representativeness and diversity within the dataset. Targets were binary-coded to distinguish between selection and neutral scenarios, enhancing the predictive models' learning efficacy. Neutral selection was coded as 0 (true positive), while selection was represented by the chosen selection coefficient (true negative). This binary encoding was implemented using the 'to_binary' method in ImaGene.

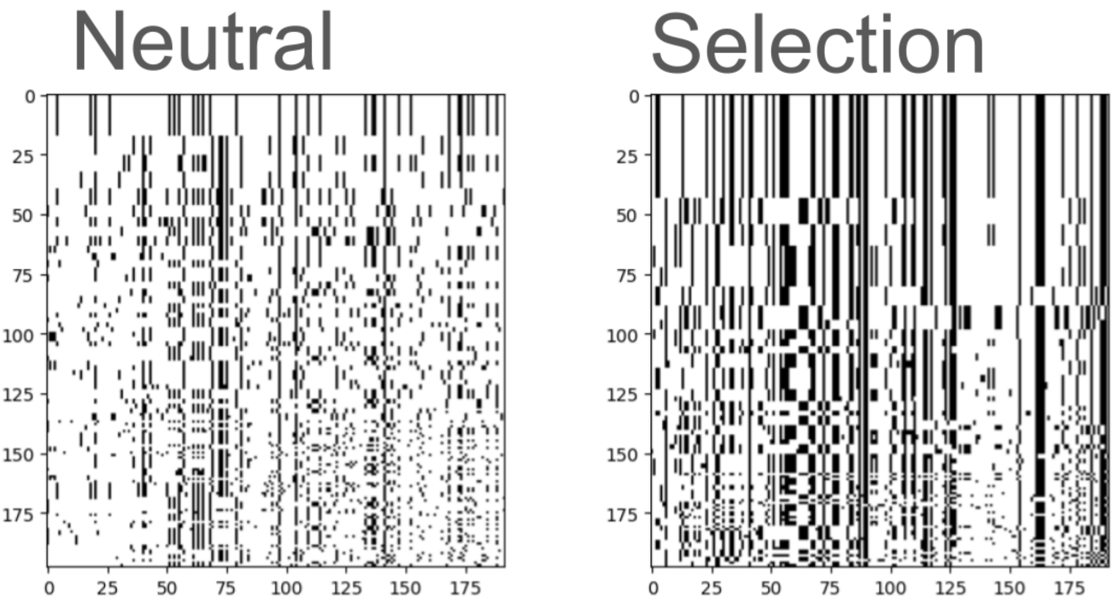


Figure 7: showing the processed Haplotype data that has undergone, resizing, sorting rows by frequency and flipping, processing steps identical to Torada et al. [13] but the resize is done to 198×192 instead of 128×128.

5.4 Training and testing procedure

The generated datasets were partitioned into training, testing, and validation sets using Python's Keras package, as depicted in the accompanying figure 8 illustrating the data split. This methodological approach ensured robust training and evaluation of the CNN. The dataset was divided into 10 batches, with batches 1-9 used for training and batch 10 reserved for model evaluation. For batches 1-9, an 80/20 training and validation split was applied to each batch, each undergoing training for 10 epochs, resulting in 90 epochs of training. The 10th batch, representing 10% of the entire dataset, was used exclusively for model evaluation. In an attempt to mitigate overfitting, a "simulation-on-the-fly" strategy was employed, akin to previous studies [13]. This iterative training method involves generating and preprocessing new simulations at each epoch, which enhances the model's generalization capability by exposing it to a continuously varying dataset, while "simulation-on-the-fly" is efficient, it does not support the replication of analysis for hyperparameter estimation [98]. The training process was conducted using google collab Jupiter notebook, specifically utilizing an NVIDIA A100 GPU and an NVIDIA L4

GPU [99]. These powerful GPUs enabled efficient processing and accelerated the training of the CNN models, ensuring timely and effective model development.

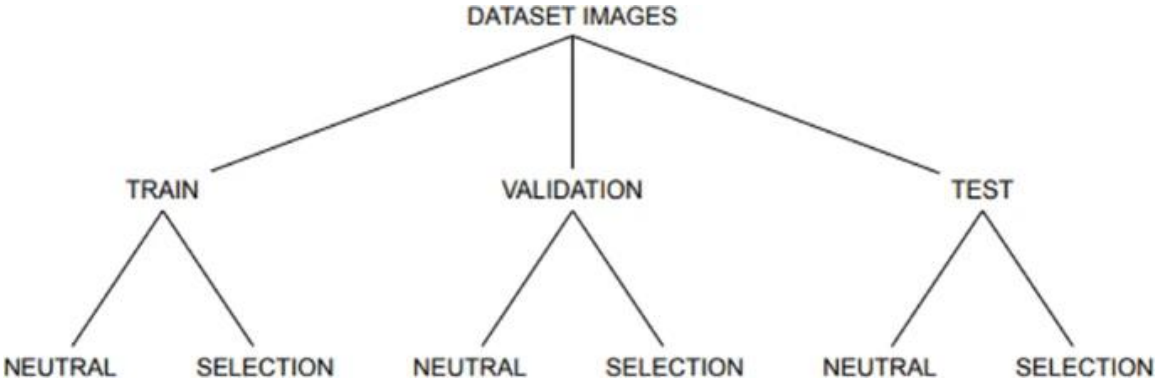


Figure 8: depicting the training and test split image taken from Fadja et al (2021) [63]

5.5 Rationale for Choosing Bayesian Optimization in Hyperparameter Tuning

In developing robust CNNs for selection inference, Bayesian optimization was employed to acquire better hyperparameters, to improve model performance. Given the numerous hyperparameter optimization (HPO) techniques available, one might question why Bayesian optimization was chosen over other approaches. This choice is informed by its demonstrated effectiveness in efficiently navigating the hyperparameter landscape. To better understand this decision, it is important to first examine and critically assess some of the commonly utilized HPO methods, including grid search, random search, and evolutionary strategies (ES).

Grid search is a traditional technique that involves discretizing the hyperparameter ranges and exhaustively evaluating all possible combinations. Although this approach is conceptually simple, it is hampered by the "curse of dimensionality," where the number of evaluations required increases exponentially as the number of hyperparameters in the search space grows, making it impractical for complex models [100]. For example, in situations where only a single hyperparameter significantly influences performance, many of the evaluations performed by grid search may prove redundant, highlighting its inefficiency in such contexts [101]. Although grid search is a model-agnostic method for finding the best hyperparameters, this also becomes its limitation. It exhaustively tests all hyperparameter configurations, including those that are unfavorable to model performance, without discriminating against poor configurations, potentially leading to inefficiencies and unnecessary computational cost and time. Moreover, Bayesian optimization outcompetes Grid search by a significant margin, in regards to the time spent in finding the optimal HPO's [125] (see Figure 9).

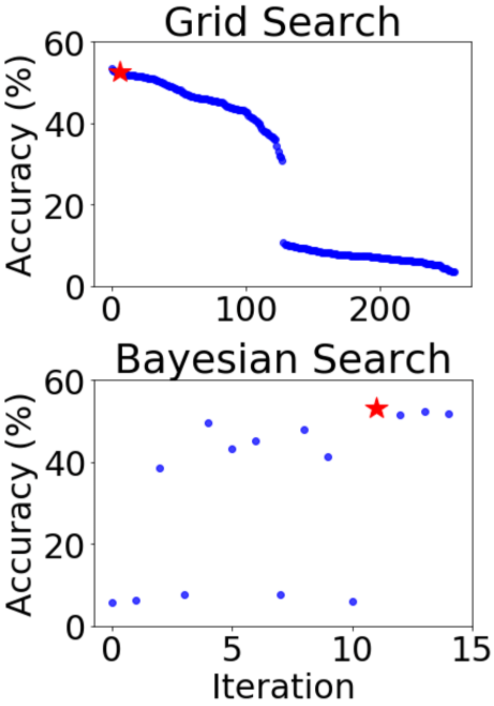


Figure 9: Bayesian optimization achieving optimal hyperparameters in significantly fewer epochs compared to grid search, highlighting Bayesian optimization's computational efficiency in intelligently navigating the hyperparameter space. Adapted from Parsa et al. (2020) [125].

Random search offers a more flexible alternative by sampling hyperparameter values independently from specified distributions, often uniformly. This method is generally more efficient than grid search, especially in high-dimensional spaces, as it can explore a wider array of configurations within the same computational budget [101]. Additionally, random search can be enhanced through techniques like Latin Hypercube Sampling [102] or Sobol Sequence [103], which aim to cover the search space more comprehensively. However random search HPO techniques still don't outperform Bayesian optimization in finding the best HPO's in the most efficient manner [104].

Evolutionary strategies (ES), inspired by the principles of natural selection, are stochastic, population-based optimization methods well-suited for black-box problems like HPO, and contextually would be more fitting for the project given its connection to evolutionary concepts. ES typically involves iterative processes such as sampling an initial population, evaluating their performance, selecting the most successful candidates, and generating new candidate solutions through mutation and crossover [105]. Although ES is effective in exploring diverse search spaces and is less likely to be trapped in local optima, it often requires many iterations to converge, resulting in significant computational overhead [106].

Having outlined the limitations of these prevalent HPO methods, it becomes evident why Bayesian optimization was selected for this study. Bayesian optimization excels by intelligently balancing the exploration of new regions and the exploitation of known promising areas, using a probabilistic model to focus on the most likely optimal hyperparameter configurations [107][108]. This capability is particularly valuable when optimizing complex or resource-intensive models. Furthermore, Bayesian optimization is versatile, accommodating a variety of hyperparameter types, including continuous, discrete, and categorical variables [109].

5.6 The Optimisation Framework

The primary objective of the hyperparameter optimization was to maximize the validation accuracy of the CNN models. This process involved systematically refining various hyperparameter configurations within a predefined hyperparameter space, conducted over 30 trials. To mitigate the risk of overfitting, an early stopping mechanism was employed, with a patience parameter set to 3. This means that the training was halted if the validation performance did not improve over three consecutive epochs, thereby enhancing the model's ability to predict accurately on unobserved data [110]. To maintain computational feasibility, the optimization training process was run for 1 epoch per batch for batches 1 through 9 of the training data, resulting in a total of 9 epochs.

The Bayesian optimization framework was managed using Optuna, a state-of-the-art hyperparameter optimization library [111]. Which leverages the Tree-structured Parzen Estimator (TPE) to efficiently navigate the hyperparameter space. The TPE method focuses on promising regions of the hyperparameter space and prunes underperforming trials early, which accelerates the optimization process and ensures computational resources are allocated to the most promising configurations [111]. The objective function in this study was defined to maximize validation accuracy. Moreover, compared to other widely used optimization libraries, such as HyperOpt, Optuna offers a superior balance between performance and computational efficiency [112] hence the justification for its incorporation in the study.

5.7 Building The Hyperparameter Space

The baseline CNN architecture provided by Imagene was initially tested across nine different datasets to evaluate its ability to learn effectively from the genomic data. However, the model quickly exhibited signs of underfitting (see figure 10), which meant that it was struggling to capture the intricate patterns within the data. Underfitting occurs when a model is too simplistic and has high bias, preventing it from capturing important patterns in the training data. This results in a high error on the training set and poor accuracy on both training and validation datasets [110] often around 50% or lower. This underperformance highlighted the need for a more complex and flexible architecture—a topic that will be explored in more detail in the results section.

To address this issue, an informed HPO space was designed to enhance the model's capacity and flexibility, thereby mitigating underfitting. The original model architecture included three convolutional layers with a limited number of filters (32, 32, and 64 respectively) and strong L1 and L2 regularization (both set to 0.005), followed by two dense layers (with 128 and 1 units, respectively). This configuration was insufficient for capturing and extracting pertinent features of the selective sweep images across more complex scenarios (scenarios other that recent and strong). The adjustments to the HPO space were informed by established research and best practices in deep learning and the initial output of preliminary optimization scripts. Firstly, the filter ranges for the convolutional layers were expanded to allow for more complex feature extraction. The number of filters per layer was set to range between 32 and 128, allowing the model to capture more detailed and varied features from the input data via increasing the number of feature maps per layer.

Additionally, kernel sizes were restricted to 2x2 or 3x3 instead of larger sizes like 5x5 to prevent excessive dimensionality reduction. Preliminary hyperparameter optimization runs indicated that larger kernels (5x5 and above) frequently caused trials to be pruned due to this issue. The ranges for L1 and L2 regularization were adjusted to 1×10^{-6} to 1×10^{-2} on a logarithmic scale. Reducing the regularization strength allowed the model to learn more freely, thus improving its ability to fit the training data [113]. Recognizing the need for a more complex model, the number of convolutional layers was set to range from 3 to 6, meaning the model could have a minimum of 3 layers, providing sufficient depth to capture complex patterns. The number of units in the dense layers was also increased, ranging from 64 to 256, adding more capacity for the model to learn more complex representations of the data [114].

The dropout rate was set between 0.1 and 0.5, with the lower end ensuring that the model retains sufficient information while still benefiting from regularization [115]. Finally, the learning rate was kept between 1×10^{-5} and 1×10^{-3} to ensure effective optimization as a proper learning rate is crucial for the convergence of the model during training [116]. Essentially, this means that the learning rate needs to be carefully set to ensure the model learns efficiently without overshooting optimal solutions or converging too slowly [116]. These adjustments to the HPO space were designed to enhance the model's ability to learn complex patterns from the data, thereby reducing underfitting and leading to better training and validation accuracy.

Hyperparameter	Type	Range/Values
num_layers	Integer	3 to 6
filters	Integer	32 to 128
kernel_size	Integer	2 to 3
l1	Float (Log scale)	1e-6 to 1e-2
l2	Float (Log scale)	1e-6 to 1e-2
dense_units	Integer	64 to 256
dropout_rate	Float	0.1 to 0.5
learning_rate	Float (Log scale)	1e-5 to 1e-3

Table 2 outlining the hyperparameter space in succinctly.

5.8 Identification, Extraction, and Preprocessing of SNPs for Genomic Analysis

The objective of this study was to also evaluate the ability of trained models to infer selection from real genomic data. To achieve this, I selected a set of genes associated with human pigmentation, each with well denoted SNPs and documented selection pressures. The pigmentation-related genes included OCA2 (rs12913832, chromosome 15), SLC45A2 (rs1426654, chromosome 15), IRF4 (rs12203592, chromosome 6), KITLG (rs12821256, chromosome 12), and TYR (rs1042602, chromosome 11) [117]. These genes, extensively studied within European populations, exhibit strong, moderate and weak selection pressures [117], making them ideal candidates for testing the models' capacity to detect selection. In addition to these pigmentation-related genes, the LCT gene was included as a control. The LCT gene, which is under strong selection in European populations and associated with lactase persistence, contains the well-documented SNP rs4988235 [118]. Data for this gene was provided by my supervisor, Matteo Fumagalli, and served as a benchmark to validate the models' ability to detect selection.

Beyond validating the models with pigmentation-related genes known to be under positive selection, this study also aimed to explore whether pathogenic SNPs within autophagy-related genes were under positive selection. The autophagy-related genes tested included ATG13 (rs4565870, rs10838611 chromosome 11), FIP200 (rs1129660 chromosome 8), ULK1 (rs9652059 chromosome 12), and ULK2 (rs281366 chromosome 17) [80]. These SNPs are associated with various diseases, such as Crohn's disease, tuberculosis, ankylosing spondylitis, and asparaginase-associated pancreatitis, underscoring their clinical significance [80]. Investigating the presence of positive selection among these pathogenic SNPs could provide insights into the evolutionary pressures that may have maintained these variants, despite their associations with disease.

SNPs and their associated rsIDs were identified via scouring the literature, once obtained the rsIDs were entered into the Ensembl genome browser, where their corresponding genomic coordinates were obtained. To ensure comprehensive genomic coverage, and to obtain data that matches the conditions of the simulated data, ±40kb on either side of the SNP was extended (from the genomic coordinates), centering it within the extracted region. These regions were then processed using the Ensembl Data Slicer to generate Variant Call Format (VCF) files, which were subsequently unzipped for further analysis. The preprocessing of the genomic data was essential to ensure compatibility with the trained models. The data was stored in an ImaFile object, where the VCF file name and the number of samples (twice the number of individuals for diploid organisms) were specified. This step was crucial to verify that the VCF file contained the expected data. The ImaFile object was then used to convert the data into ImaGene objects, where a series of standardized filtering and transformation steps were applied, which was an identical process to the preprocessing of the simulations.

6.0 Results

Confusion matrices and key performance metrics, including F1 Score, Precision, Recall, and AUC, were incorporated to evaluate the model's performance across various evolutionary scenarios. ROC curves and confusion matrices were displayed as a matrix of matrices to assess the architecture’s performance under different evolutionary scenarios before and after optimization, with the selection coefficient on the y-axis and selection time on the x-axis.

In this context:

- A true positive is correctly classifying neutral selection as neutral selection.
- A true negative is correctly classifying positive selection as positive selection.

- A false positive occurs when positive selection is incorrectly classified as neutral selection.
- A false negative occurs when neutral selection is incorrectly classified as positive selection.

Ideal performance is indicated by values close to or exactly 1 in "**both"** the top-left and bottom-right diagonal boxes of the confusion matrix. For this section, results will be referred to using shorthand terminology such as true positive rate, false positive rate, etc. Moreover, ROC curve scores well above 0.50 (the ROC threshold) signify strong discriminative power [119] indicating a good binary classifier. After training, the models were loaded and used to predict whether a genomic region (an 80kb SNP, with the variant of interest centered) was under positive selection across the different evolutionary scenarios. This was visualized as a matrix of bar charts, where the x-axis represented the type of selection on which the models were trained, and the y-axis displayed the predicted class label probabilities whereby higher scores indicate increased chance of being under that type of positive selection. Additionally, ATG13_S and ATG13_B are designations used for clarity in this study to distinguish between the ATG13 gene variant associated with Selective IgA Deficiency (ATG13_S) and the variant associated with Breast Cancer (ATG13_B).

6.1 Low complexity architecture against evolutionary scenarios

6.1.1 Recent Selective Sweeps (10kya)

Weak (S = 100), and moderate (S = 200) selection pressures demonstrated significantly poor performance in predictive capability, predicting all instances as neutral selection. The overall accuracy of the models for both weak and moderate selection remained at 50% (see figure 10), which elucidates the predictive power to be akin to random chance guessing, indicating that the model failed to learn any meaningful patterns in the dataset. The F1 Score, precision, and recall all registered at 0.0 for both these evolutionary scenarios (see Table 3) and a ROC curve area of 0.50 indicating no discriminative power (see figure 11), further underscoring the model's inability to detect subtle signals of weak-to-moderate selective pressure. In contrast for strong selection (S=300), the model's performance improved drastically. It achieved an almost perfect overall accuracy of 99.70%, with the confusion matrixes highlighting a true positive rate of 99.75% and a true negative rate of 99.65% (see figure 10) with minimal misclassification, that being a less than 1% false negative and false positive rate, and a ROC Curve Area of 1.0 (see figure 11) with near perfect F1, precision and recall (0.997, 0.998, 0.997, see Table 3). This substantial improvement underscores how low complexity models effortlessly detect recent and strong selective sweeps which has been observed in the literature [60][13].

6.1.2 Intermediate Selective Sweeps (50kya)

Under weak selection (S = 100), the model achieved an overall accuracy of 85.02% (see figure 10). The confusion matrix revealed that the model had a true positive rate of 85.25% and a true negative rate of 84.80%, with an F1 Score of 0.850, Precision of 0.851, and recall of 0.848 (see table 3). The ROC curve Area was 0.93 (see figure 11). These results suggest that while the model performs reasonably well in detecting weak selection signals 50kya, a notable portion of instances are misclassified as the model had a false positive rate of 14.75% and a false negative rate of 15.20%, likely due to the older timing of selection attenuating the selection signal.

However, under moderate selection (S = 200), the model's accuracy dramatically dropped to 50% (see figure 10), with all instances classified as selection. This resulted in an F1 Score of 0.667, precision of 0.5, recall of 1.0 (see table 3), and an ROC Curve Area of 0.50 (see figure 11). The model's strong bias toward predicting selection in all cases indicates a potential issue of class imbalance, leading to an inability to generalize effectively. Notably, these same results were observed under strong selection (S = 300), where the model continued to misclassify, achieving only 50% accuracy. It failed to correctly identify any instances of neutral selection correctly and identical to the model trained under moderate selection classified all the instances as positive selection. This consistent performance across both moderate and strong selection scenarios underscores a significant limitation in the low-complexity model's ability to accurately detect and differentiate moderate-to-strong selection signals at this intermediate temporal distance.

6.1.3 Ancient Selective Sweeps (100kya)

Under weak selection (S = 100), the model achieved an overall accuracy of 76.05% (see figure 10). The confusion matrix revealed a true positive rate of 96.20% and a true negative rate of 55.90%. However, this performance was accompanied by a notable false negative rate of 44.10% and a false positive rate of 3.80%. The F1 Score stood at 0.700, with a precision of 0.936 and recall of 0.559 (see table 3), and an ROC Curve Area of 0.81 (see figure 11). These results indicate that while the model exhibited moderate success in detecting weak selection signals, its sensitivity to such signals was significantly compromised over this extended temporal distance. As the model evidently struggles to correctly classify positive selection nearing a true negative accuracy of 50%.

Under moderate selection (S = 200), the model’s performance drastically declined, with accuracy falling to 50% (see figure10). All instances were incorrectly classified as neutral selection leading to an F1 Score, precision, and recall all registering at 0.0 (see table 3), and an ROC Curve Area of 0.50 (see figure 11). This pronounced bias towards neutral selection suggests that the selection signal at this ancient temporal distance is largely undetectable by the model, resulting in complete misclassification and essentially displaying no discriminative power.

Similarly, under strong selection (S = 300), the model's performance remained suboptimal, again with an accuracy of 50% (see figure 10), but this time classifying all instances as selection. This indicates a persistent challenge for the model in distinguishing strong selection signals that have persisted over long timescales, indicating no discriminative power.

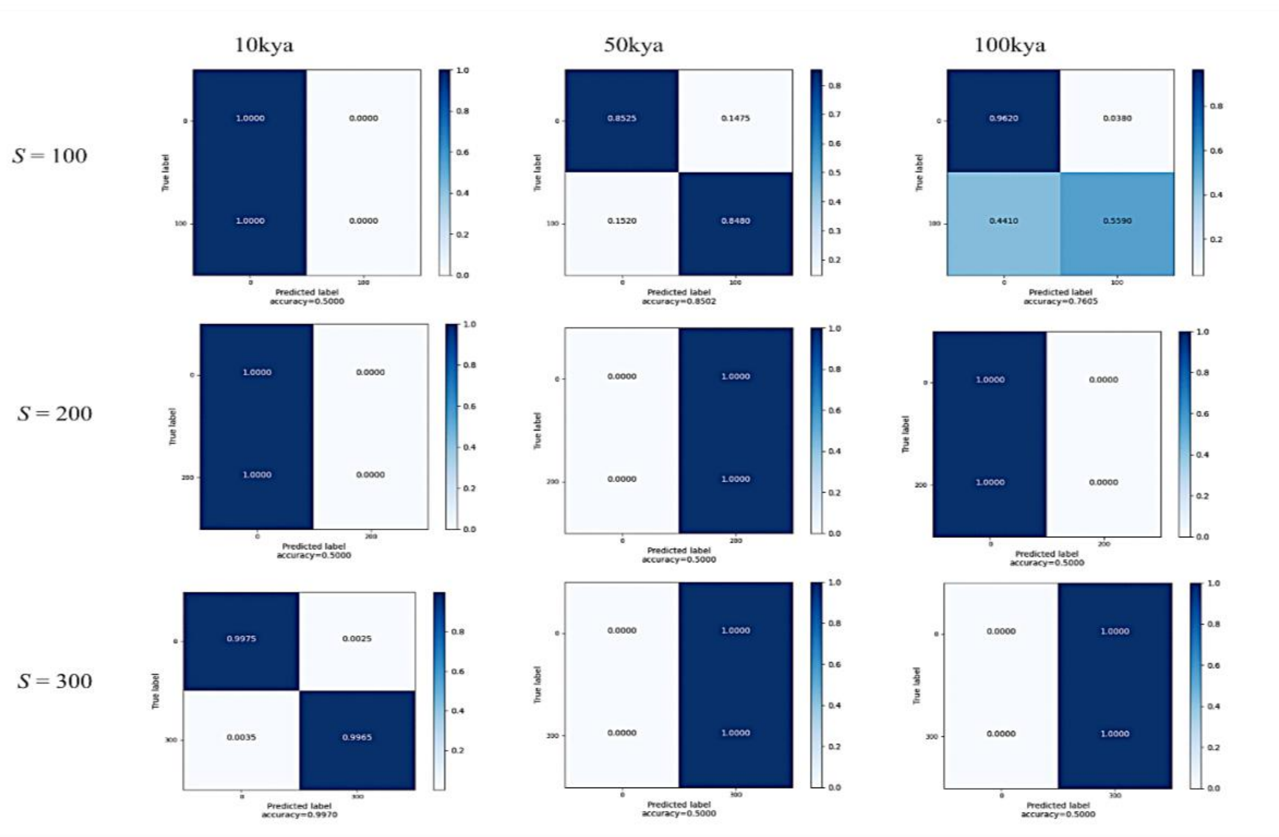


Figure 10: Confusion matrices showing the accuracy of detecting positive selection across different temporal distances (10kya, 50kya, 100kya) and selection coefficients (S = 100, S = 200, S = 300). Neutrality is represented as 0, and selection corresponds to the specified selection coefficient (S). These results are based on the 3-layer baseline ImaGene CNN architecture, illustrating the model's performance in classifying genomic regions as either neutral (N) or under selection (S).

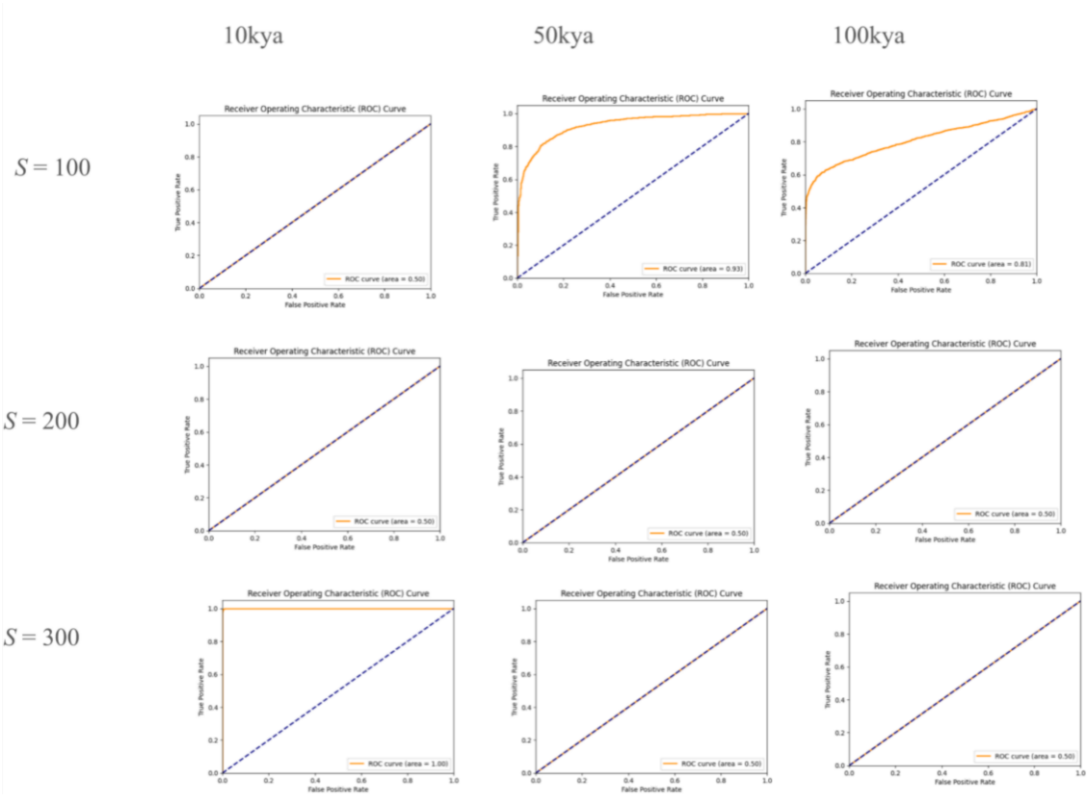


Figure 11: Overview of ROC Curves Displays ROC curves for the 3-layer baseline ImaGene CNN across different temporal distances (10kya, 50kya, 100kya) and selection strengths (S = 100, S = 200, S = 300), with rows and columns organized by selection strength and temporal distance, respectively. Each curve features a reference line at 0.50 to indicate performance above random chance. The AUC values highlight low discriminative power in distinguishing selection from neutrality, notably declining with reduced selection strength and increased temporal distance, except in ancient and weak, and intermediate and weak selections, which perform well.

Table 3: Performance metrics across temporal scales and selection strengths

	Key Metrics		
	F1 Score	Precision	Recall
Recent strong	0.997	0.998	0.997
Recent moderate	0.0	0.0	0.0
Recent weak	0.0	0.0	0.0
Intermediate strong	0.667	0.5	1.0
Intermediate moderate	0.667	0.5	1.0
Intermediate weak	0.850	0.852	0.849
Ancient strong	0.667	0.5	1.0

Ancient moderate	0.0	0.0	0.0
Ancient weak	0.700	0.936	0.559

Note. F1 Score, Precision, and Recall values are presented for each combination of temporal scale and selection strength for the baseline ImaGene model

6.2 Bayesian Optimized Architecture and Evolutionary Scenarios

This segment compares the performance of models post-Bayesian optimization. The optimization explored a broad hyperparameter space, resulting in improved performance across various evolutionary scenarios. Optimization for recent and strong selection was omitted due to the model already achieving near-optimal performance on the low-complexity model, rendering further optimization redundant.

6.2.1 Recent Selective Sweeps

Weak selection (S = 100) resulted in the model achieving an overall accuracy of 50.40%. The confusion matrix revealed a true positive rate of 48.95% and a true negative rate of 51.86%, alongside a false negative rate of 48.15% and a false positive rate of 51.85% (see figure 12). The model's F1 Score was 0.5111, with a Precision of 0.5039 and a Recall (Sensitivity) of 0.5185 (See table 4). The Area Under the ROC Curve (AUC) was 0.50 identical to the baseline performance but upon visual inspection the ROC curve was observed to display mild Stochasticity above and below the 0.50 threshold line, suggested very weak discriminative power (see figure 13). The model's performance showed only a marginal improvement over random chance and a slight enhancement compared to the baseline architecture, indicating limited benefits from the hyperparameters elucidated from Bayesian optimization most likely due to how difficult it to detect weak and recent selective sweeps as the signals hasn’t had enough time to create a distinct signal from neutrality.

In contrast, moderate selection (S = 200) saw a substantial improvement in model performance compared to the baseline architecture, with an overall accuracy of 84.55%. The confusion matrix showed a true positive rate of 90% and a true negative rate of 79.10%, with a false negative rate of 20.90% and a false positive rate of 10% (See figure 12)

The model's F1 Score increased to 0.837, with a Precision of 0.888 and a Recall of 0.791(see table 4). The ROC curve was 0.92 (see figure 13), highlighting a marked improvement compared to the baseline model, with a strong capability to differentiate between instances of positive selection and instances of neutral selection at this selection strength and temporal distance.

6.2.2 Intermediate Selective Sweeps

For Weak selection (S = 100), the model achieved an overall accuracy of 88.88%. The confusion matrix revealed a true positive rate of 96.50% and a true negative rate of 81.25%, alongside a false negative rate of 18.75% and a false positive rate of 3.50% (see figure 12). The F1 Score was 0.879, with a Precision of 0.959 and a Recall of 0.813(see table 4). The Area Under the ROC Curve was 0.96 (see figure 13), indicating near excellent discriminative power.

For moderate selection (S = 200), the model's overall accuracy was 82.87%. The confusion matrix indicated a true positive rate of 89.60% and a true negative rate of 76.16%, with a false negative rate of 23.85% and a false positive rate of 10.40% (see figure 12). The F1 Score was 0.816, with a Precision of 0.879 and a Recall of 0.762 (See Table 4). The roc curve value was 0.91(see figure 13), reflecting strong but slightly diminished performance compared to the weaker selection scenario but showed a substantial increase over the baseline architecture showing strong discriminative power.

For strong selection (S = 300), the model's performance further declined, with an overall accuracy of 79.57%. The confusion matrix showed a true positive rate of 78.85% and a true negative rate of 80.30%, with a false negative rate of 19.70% and a false positive rate of 21.15% (see figure 12). The F1 Score was 0.797, with a Precision of 0.792 and a Recall of 0.803 (see table 4). The AUC was 0.88 (see figure 13), indicating a noticeable increase in the discriminative power of the model compared to the baseline.

6.2.3 Ancient Selective Sweeps

For weak selection (S = 100), the model achieved an overall accuracy of 77.60%. The confusion matrix revealed a true positive rate of 85.85% and a true negative rate of 69.35%, alongside a false negative rate of 30.65% and a false positive rate of 14.15% (see figure 12). The F1 Score was 0.756, with a Precision of 0.831 and a Recall of 0.694 (see table 4). The Area Under the ROC curve was 0.84 (see figure 13), indicating reasonably strong discriminative power of the model under weak selection and very strong discriminative power. However, only a minor increase in performance compared to the baseline architecture.

For moderate selection (S = 200), the model's overall accuracy decreased to 69.10%. The confusion matrix indicated a true positive rate of 88% and a true negative rate of 50.20%, with a false negative rate of 49.80% and a false positive rate of 12% (see figure 12). The F1 Score was 0.6190, with a Precision of 0.8071 and a Recall of 0.502 (see table 4). The ROC curve value was 0.75 (see figure 13), reflecting a decline in the model's discriminative power compared to weak and ancient selection, but it highlighted a significant increase compared to the baseline performance which had no discriminative power.

Under strong selection (S = 300), the model demonstrated improved performance compared to baseline and compared to the other selection scenarios within the same temporal distance of weaker selection coefficients, achieving an overall accuracy of 88.42%. The confusion matrix showed a true positive rate of 91.35% and a true negative rate of 88.50%, with a false negative rate of 14.50% and a false positive rate of 8.65% (see figure 12). The F1 Score was 0.881, with a Precision of 0.908 and a Recall of 0.855(see table 4). The ROC was 0.94 (see figure 13), indicating a high level of accuracy and a robust capability to differentiate between selection and no selection in this evolutionary context highlighting significant improvement compared to the low complexity, baseline model provided by ImaGene.

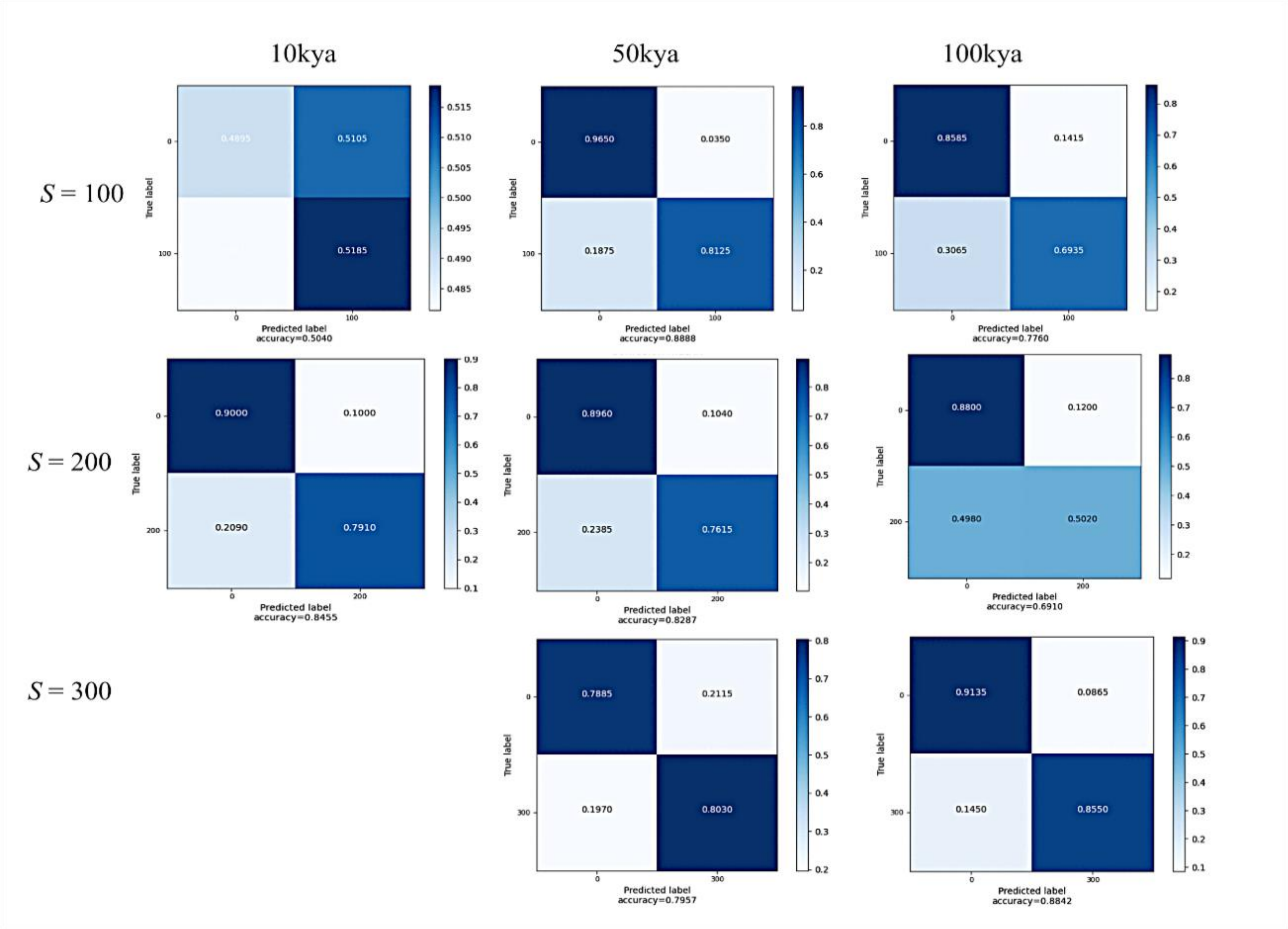


Figure 12: Confusion matrices showing the accuracy of detecting positive selection across different temporal distances (10kya, 50kya, 100kya) and selection coefficients (S = 100, S = 200, S = 300). Neutrality is represented as 0, while selection corresponds to the specified selection coefficient (S). These results are based on the Bayesian optimized ImaGene CNN architectures, illustrating the enhanced performance in classifying genomic regions as either neutral (N) or under selection (S), Recent and strong selection has been omitted due to achieving near optimal performance at baseline

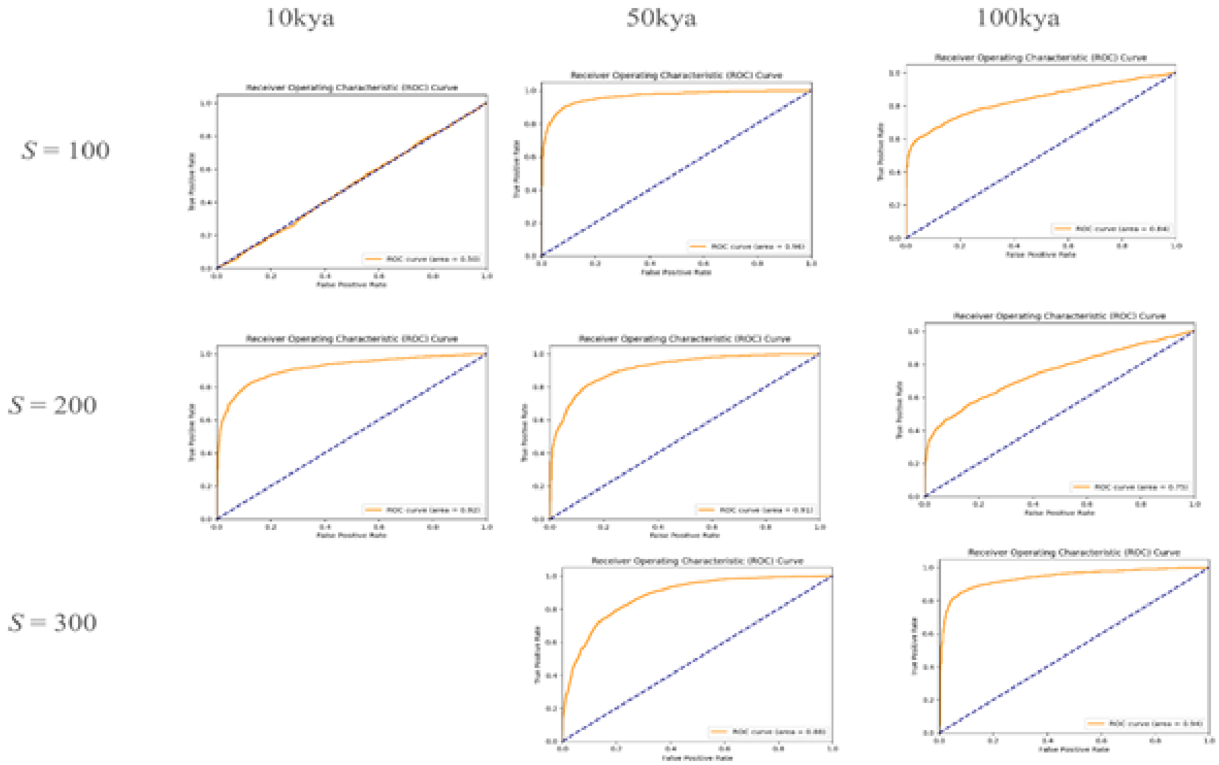


Figure 13: Overview of ROC Curves Displays ROC curves for the baseline ImaGene CNN across different temporal distances (10kya, 50kya, 100kya) and selection strengths (S = 100, S = 200, S = 300), with rows and columns organized by selection strength and temporal distance, respectively. Each curve features a reference line at 0.50 to indicate performance above random chance. The AUC values highlight effectiveness in distinguishing selection from neutrality across all evolutionary scenarios

Table 4: Performance metrics across temporal scales and selection strengths

	Key metrics		
	F1 Score	Precision	Recall
Recent moderate	0.837	0.888	0.791
Recent weak	0.511	0.503	0.519
Intermediate strong	0.797	0.792	0.803
Intermediate moderate	0.816	0.879	0.762
Intermediate weak	0.879	0.959	0.813
Ancient strong	0.881	0.908	0.855
Ancient moderate	0.619	0.807	0.502
Ancient weak	0.756	0.831	0.694

Note. F1 Score, Precision, and Recall values are presented for each combination of temporal scale and selection strength for the Bayesian Optimised ImaGene model.

6.3 Architectural patterns identified using Bayesian optimization

Table 5 below summarizes the architectural patterns that were discovered through Bayesian optimization across the evolutionary scenarios except recent and strong selection, compared to the baseline ImaGene model. Key features of each architecture, along with their computational complexity are presented. FLOPs were calculated using TensorFlow tools such as the TensorFlow Profiler and serve as indicators of algorithmic complexity and proxies for neural network speed [129] (More detail can be seen in Appendix A table 6)

Selection Scenario	Key Features	Computational Complexity vs Baseline
Recent and Weak	4 convolutional layers, varying filter sizes (39-101), smaller kernel sizes (2x2), reduced regularization, 238 dense units, dropout layer	1.168x more FLOPs with fewer parameters
Recent and Moderate	5 convolutional layers, reduced L1 and L2 regularization, 490,931 parameters, 8x more FLOPs	8x more FLOPs with fewer parameters (490,931 vs. 4,173,473)
Intermediate and Weak	6 convolutional layers, smaller kernel sizes (2x2), diverse filter sizes, reduced L1 and L2 regularization, dropout layer, 203,971 parameters	1.3x more FLOPs with fewer parameters
Intermediate and Moderate	3 convolutional layers, larger filter numbers (79-83), smaller kernel size (2x2), reduced L1 regularization, increased L2 regularization, dropout layer, over 5 million parameters	2.45x more FLOPs more parameters
Intermediate and Strong	4 convolutional layers, varying filter sizes (37-124), reduced L1 and minimal L2 regularization, dropout layer, 205 dense units	9.5x more FLOPs with fewer parameters
Ancient and Weak	3 convolutional layers, more filters per layer (79, 79, 125), smaller kernel sizes (2x2), reduced L1 and increased L2 regularization, 227 dense units, dropout layer	2.7x more FLOPs more parameters
Ancient and Moderate	4 convolutional layers, varied filter numbers (50, 91, 73, 118), smaller kernel sizes (2x2), higher L1 and lower L2 regularization, 142 dense units, dropout layer	2x more FLOPs less parameters
Ancient and Strong	4 convolutional layers, filter counts (88, 38, 88, 106), 3x3 kernel size, lower L1 and higher L2	3x more FLOPs less parameters.

	regularization, 193 dense units, dropout layer	

Table 5. shows the computational changes in terms of flops and parameters that arose due to Bayesian optimization.

6.4 Model Predictions On Known Candidate Genes

This section of the experimental design details the application phase, during which the trained models were utilized to predict known variants under positive selection with the aim of inferring the timing and to confirm the strength. The approach was specifically tested on human pigmentation genes, known to be under positive selection, as well as on genes associated with lactose persistence. All the results are presented as bar charts in figure 14.

The variant rs4988235 within the LCT gene has been extensively documented in the literature as being under recent and strong positive selection [118][120]. The gene consistently exhibited high probabilities across different time periods: 0.9947 for ancient strong selection, 0.9130 for ancient moderate selection, and 0.9989 for ancient weak selection; 0.9834 for intermediate strong selection, 0.9999 for intermediate moderate selection, and 0.9993 for intermediate weak selection; and 0.9998 for recent strong selection and 0.9029 for recent moderate selection with the highest values being in the 10-50kya time-scales corroborating with the literature that LCT gene has been under selection in shorter rather than longer temporal distances. Also, these results the validate the models' effectiveness in detecting evolutionary pressures in genomic data, marking a significant success of the project.

Moving on, the variant in the SLC45A2 (rs1426654, chromosome 15) gene known to be under strong selection [117] also showed consistently high probabilities across most selection scenarios, indicating sustained selection pressures over time. The highest probability was for the ancient strong selection model (0.9999975), suggesting selection in ancient times. Other ancient selection models also had high probabilities (0.9963 for moderate and 0.99997 for weak selection). During the intermediate period, strong selection persisted, with probabilities of 0.9438 for strong, 0.9999 for moderate, and 0.9999 for weak selection. In recent periods, while the probabilities are slightly lower (0.9235 for strong and 0.6096 for moderate selection), they still suggest ongoing selection, particularly related to the evolution of lighter skin pigmentation in European populations. The alignment of these results with the literature further confirms the robustness of the model’s predictions and uncovers the high probability that the gene is under ancient selection and strong selection.

The variant in the TYR gene (rs1042602, chromosome 11) known to be under moderate selection [117] exhibited varied probabilities across evolutionary scenarios, with the highest observed under intermediate moderate selection (0.9999825), indicating significant selective pressure during this time period. High probabilities were also noted for intermediate weak selection (0.9993836) and recent strong selection (0.9996603), suggesting continued but less intense selection more recently. In ancient scenarios, probabilities were generally lower, particularly for ancient strong selection (0.0313), indicating weaker selection pressure. However, recent moderate selection still showed a high probability (0.997465), suggesting ongoing, though diminished, selection pressure on TYR. The alignment of the model predictions with established literature validates its accuracy with models suggesting that the variant was selected for at an intermediate temporal distance.

The variant in the OCA2 gene (rs12913832, chromosome 15) known to be under strong selection pressure [117], displayed consistently high-class label values across various selection scenarios, indicating persistent and strong selection pressures throughout different evolutionary periods. In the ancient selection scenarios, the gene exhibited probabilities of 0.91677576 for ancient and strong selection, 0.8746481 for ancient and moderate selection, and 0.9988042 for ancient and weak selection. During the intermediate period, the gene displayed even higher probabilities, with the highest value being for intermediate and moderate selection (0.99978894, this was the highest value across all time scales), other values for selection 50kya were intermediate and strong which scored 0.99978894. Additionally, the probability was 0.9969222 for intermediate and weak selection. In recent times, the gene continued to show high probabilities, with values of 0.91350234 for recent strong selection and 0.9834473 for recent moderate selection. These results suggest that the OCA2 gene remains under considerable selection pressure in contemporary populations. The results could be argued to align with the literature showing moderate-to-strong selection pressure, as the highest values were observed in the intermediate and moderate selection category and very high values were observed for strong selection pressure also.

The variant in the gene KITLG (rs12821256, chromosome 12), known to be under weak selection [117] demonstrated consistently high probabilities across various selection scenarios. In the ancient period, probabilities were 0.9705227 for strong selection, 0.9340623 for moderate selection, and 0.99936634 for weak selection. During the intermediate period, the gene showed a wide range of probabilities, from 0.41052908 for strong selection to 0.99999094 for moderate selection, which was the highest probability observed across all scenarios, and 0.99901533 for weak selection. In recent times, KITLG continued to display high probabilities, with 0.99997246 for strong selection and 0.9919308 for moderate selection. These results slightly deviate from the literature; however, the highest values being associated with intermediate and moderate selection, alongside ancient and weak, as well as intermediate and weak selection, coupled with the sharp decrease in the class label probability for intermediate strong selection, suggest that KITLG is likely under weak or moderate-to-weak selection, as the findings indicate. Moreover, the intermediate timing of selection could also be said to be corroborated by the literature, as KITLG has been reported to have been selected 30,000 years ago, which could be deemed intermediate.

Lastly, the variant in the gene IRF4 (rs12203592, chromosome 6) known to be under weak selection displayed relatively low-class label values across most selection scenarios, suggesting weaker selection pressures throughout its evolutionary history. In the ancient selection scenarios, the gene showed modest probabilities, with 0.16579019 for ancient and strong selection, 0.225036 for ancient and moderate selection, and 0.37499258 for ancient and weak selection. These values indicate that the IRF4 gene was likely under limited selection pressure during ancient times. During the intermediate period, the probabilities varied, with the highest value of 0.89342886 observed for intermediate and moderate selection, suggesting a strong degree of selection approximately 50,000 years ago under a moderate selection coefficient. In contrast, the probabilities for intermediate and strong selection (0.1745862) and intermediate and weak selection (0.007158009) were much lower, indicating that the selection pressure during this period was primarily moderate. The selection in more recent periods was negligible, with probabilities for recent and weak selection at 0.020208418 and recent and moderate selection at 0.0008955849. The highest class label value being for intermediate and moderate selection, alongside ancient and weak selection, corroborates that IRF4 has been under weak-to-moderate selection, consistent with the literature.

Genes under selective pressure.

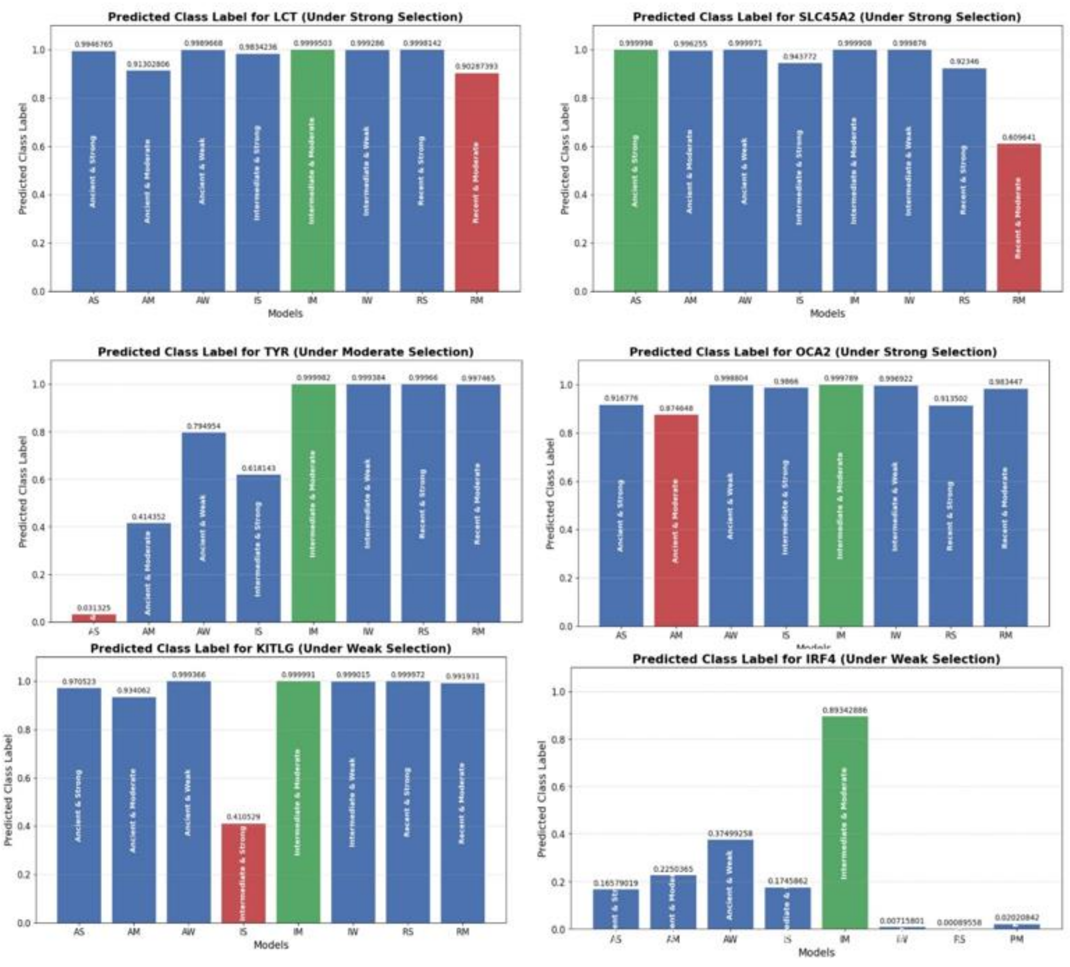

Predicted Class Label for LCT (Under Strong Selection)
Predicted Class Label for SLC45A2 (Under Strong Selection)
Predicted Class Label for TYR (Under Moderate Selection)
Predicted Class Label for OCA2 (Under Strong Selection)
Predicted Class Label for KITLG (Under Weak Selection)
Predicted Class Label for IRF4 (Under Weak Selection)

Figure 14: Predicted class labels for six genomic regions. Five associated with human pigmentation (SLC45A2, TYR, OCA2, KITLG, IRF4) and one control (LCT, linked to lactose persistence)—are shown across different selection models. The x-axis represents models trained on categories of selection (e.g., ancient, and strong, recent, and moderate), while the y-axis indicates the predicted class label. Higher class label values suggest a greater likelihood that the genomic region is under the type of selection corresponding to the model.

6.5 Model Predictions On Autophagy Dysregulation Genes

This section explores the results from testing trained models on disease genes associated with autophagy dysregulation, particularly those within the ULK1 complex, which are linked to conditions such as hypertension, breast cancer, and ankylosing spondylitis [80]. This investigation is entirely exploratory due to the lack of prior hypotheses or predictions, given the untested nature of the subject.

The variant in the ATG13_S (rs4565870, chromosome 11) gene exhibited varying probabilities across different selection scenarios. The highest probability, 0.9699, was observed under intermediate and moderate selection, suggesting moderate selection pressure around 50,000 years ago. Other intermediate scenarios displayed probabilities of 0.5035 for intermediate and strong selection and 0.0807 for intermediate and weak selection, further indicating that moderate selection was predominant during this period. For ancient selection scenarios, the gene showed probabilities of 0.5366 under strong selection, 0.5371 under moderate selection, and 0.6977 under weak selection, suggesting that weak-to-moderate selection pressures were likely present during ancient times. However, in recent selection scenarios, the probabilities were significantly lower, indicating that the gene is not under strong selection in contemporary times (Recent and moderate: 0.11559053, Recent and strong: 0.0004986264). Overall, the findings suggest that the ATG13_S gene may have been subject to selection during both ancient and intermediate periods, particularly around 50,000 years ago, with a notable decline in selection pressure in more recent times.

The ATG13_B gene displayed its highest probability, 0.9996, under intermediate and moderate selection. Additional intermediate scenarios showed high probabilities, with 0.8672 under intermediate and strong selection and 0.9808 under intermediate and weak selection, suggesting consistent selection during this period. In ancient selection scenarios, the probabilities ranged from 0.8346 for ancient and strong selection to 0.4978 for ancient and moderate selection, with a prominent signal under ancient and weak selection at 0.9002. Recent selection scenarios revealed a sharp decline in moderate selection probability to 0.1425, although strong selection still showed a probability of 0.8849. These results imply that the ATG13_B gene was under persistent selection pressure around 50,000 years ago, with a reduction in selection intensity in recent times.

Similarly, the FIP200 gene, which is associated with hypertension, showed the highest probability of 0.9872 under intermediate and moderate selection, pointing to significant selection pressure around 50,000 years ago. Other intermediate scenarios also exhibited high probabilities, reinforcing the indication of substantial selection during this period. In ancient selection scenarios, the strongest signal was observed under weak selection with a probability of 0.9467, while the probabilities for strong and moderate selection were 0.8469 and 0.5076, respectively. Recent selection scenarios, however, revealed much lower probabilities, with 0.0001 under strong selection and 0.1649 under moderate selection, indicating that the FIP200 gene is no longer under significant selection pressure in modern times.

The ULK1 gene, which is linked to ankylosing spondylitis, also showed high probabilities in intermediate selection scenarios, peaking at 0.9894 under intermediate and moderate selection around 50,000 years ago. Intermediate and strong selection yielded a probability of 0.9568, while intermediate and weak selection showed 0.9542, all suggesting robust selection pressures during this time. Ancient selection scenarios exhibited slightly lower but still significant probabilities, with 0.7756 under ancient and strong selection, 0.4164 under ancient and moderate selection, and 0.8284 under ancient and weak selection. In contrast, recent scenarios indicated a decline in selective pressure, with probabilities of 0.6997 under recent and strong selection and 0.0172 under recent and moderate selection. The data suggest that the ULK1 gene likely experienced strong selection around 50,000 years ago, with diminishing pressures in more recent times.

In contrast to the other genes analysed, the ULK2 gene, associated with asparaginase-associated pancreatitis, consistently showed low probabilities across all selection scenarios, suggesting minimal selection pressure throughout its evolutionary history. The highest probabilities were observed in ancient selection scenarios, with 0.3975 under weak selection, 0.3809 under moderate selection, and 0.0326 under strong selection. Intermediate period scenarios showed even lower probabilities, with 0.0281 for intermediate and strong selection, 0.0301 for intermediate and moderate selection, and 0.0006 for intermediate and weak selection, indicating very limited selective influence around 50,000 years ago. Recent selection scenarios also revealed very low probabilities, with 0.0008 for recent and strong selection and 0.0315 for recent and moderate selection. These findings strongly suggest that the ULK2 gene has not been under significant selection pressure, consistent with the hypothesis that it may not have played a major role in evolutionary adaptation, particularly in relation to asparaginase-associated pancreatitis.

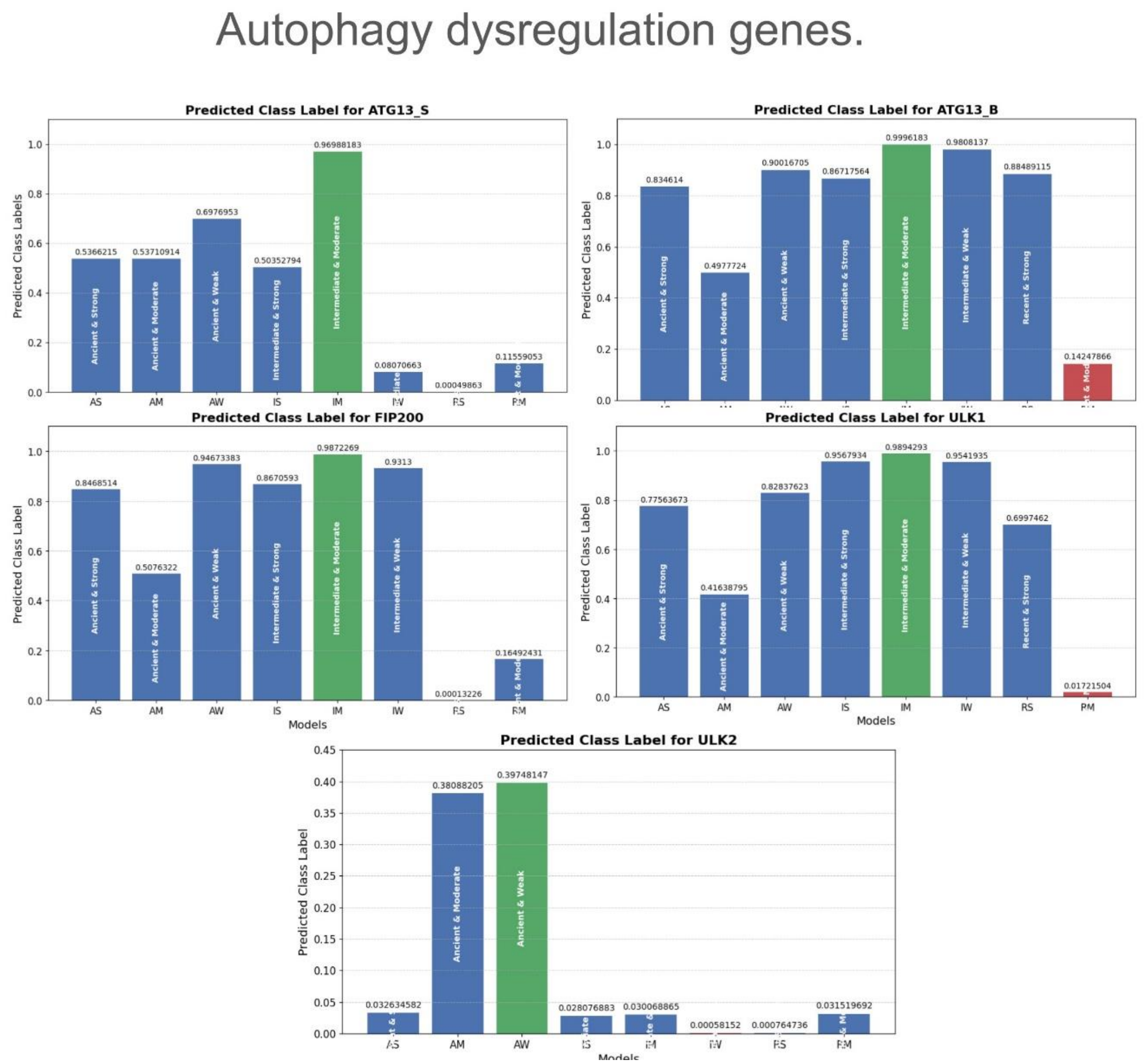


figure 15: Predicted class labels for five genomic regions, all associated with autophagy dysregulation on ULK1-complex (ULK2, FIP200, ULK1, ATG13_2, ATG13_B). The x-axis represents the models trained on the different categories of selection (e.g., ancient, and strong, recent, and moderate), while the y-axis indicates the predicted class label. Higher values suggest a greater likelihood that the genomic region is under the type of selection corresponding to the model.

7.0 Discussion

In this study, we enhanced the ImaGene framework, a parametric tool for detecting and inferring signatures of natural selection, to explore whether CNNs can identify selection signatures over ancient time scales, addressing the issue of temporal myopia often seen in the literature. We created datasets representing selection events that occurred 400, 2,000, and 4,000 generations ago, across various selection strengths (s = 100, 200, 300). In summary, the results indicate that CNNs can effectively detect ancient and nearly neutral selection scenarios

Our results show that low-complexity networks are inadequate for reliably inferring moderate-to-strong selection signals from ancient time periods (over 2000 generations ago). This difficulty in detecting ancient hard selective sweeps has been documented in the literature. For instance, Alexandre M Harris et al., 2018 [95] used an ensemble of summary statistic methods (used: G12, G123, H12, H123) and found similar challenges. Although their methods differed from those used in this study, both studies emphasize the considerable difficulty in detecting hard selective sweeps 2000 generations and beyond. This is most likely due to the increased time away from fixation attenuating the strong selection signal which gets eroded by mutations and recombination.

Moreover, low complexity models are also insufficient in detecting moderate-to-weak selection in very recent time periods (400 generations ago) this lower accuracy in being able to detect recent weak-to-moderate selection has also been seen in the literature, where Schrider et al (2016)[53] also saw a reduced ability in accuracy in detecting weak selection in recent temporal time scales compared to strong selection, albeit he observed a higher overall ROC a reduction was still observed . Other studies have also reported reduced accuracy in detecting recent weak-to-moderate selective sweeps, or a complete lack of power in identifying recent weak-to-moderate selection [121][122][60]. The difficulty in detecting weak selective sweeps in recent time periods arises because these sweeps progress slowly, needing more time to generate a distinct signal that sets them apart from neutrality [95]. This study also shows that for a weak selective sweep, more than 400 generations are required to clearly identify the signal [97].

However, it is notable to mention that Torada et al. (2019) [13] achieved high accuracy for moderate and recent selection, with accuracies exceeding 70% for selection coefficients of S = 200 across various sorting algorithms. In contrast, my analysis for the same selection coefficient (S = 200) yielded an accuracy of approximately 50%. This discrepancy may be attributed to differences in the number of simulations per dataset, as the current study used significantly fewer simulations than Torada et al., who employed over 2 million simulations, potentially leading to more robust predictive outcomes in their analysis. Also, the discrepancy could be due to his simulation-on-the-fly approach which prevents reproducible analysis or the stochastic nature of neural network random weight initialization [128].

Low-complexity networks tend to perform well only in detecting selection in more recent and strong evolutionary scenarios, typically within the last 400-800 generations which has been observed within the contemporary literature [60][13] as previous work has even identified recent and strong selective sweeps with even simpler architectures than the one provided by ImaGene [13]. This outcome is anticipated because selective sweeps with a high selection coefficient rapidly achieve fixation within the population leaving a very strong and noticeable deviation from neutrality. Once fixation occurs, mutation and recombination gradually disrupt the regions of increased homozygosity [95]. As a result, the greater the temporal distance from fixation, the more the strong selection signal degrades. Meaning over long temporal distances the strong selection signal erodes but over short temporal distances the strong selection signal is very easy to detect.

Low-complexity networks were particularly effective in detecting weak selection signals at older temporal distances, such as 2,000 generations ago. However, as these distances increased, model accuracy diminished, especially for ancient signals. For instance, the baseline architecture maintained 85.02% accuracy at 50,000 years ago, but this dropped to 76.05% at 100,000 years ago, highlighting that although these signals remain detectable, their strength fades over time—a pattern also observed in the literature [95]. As previously noted, alleles under weaker selection pressures have a slower rise to fixation compared to alleles under strong selective pressure, requiring increased temporal distance to produce a detectable signal.

7.1 Bayesian Optimization

Bayesian optimization unanimously increased the accuracy across all scenarios. It favoured wider and deeper architectures with significantly more computational complexity due to an increased FLOP count, though many of the networks required fewer parameters. Bayesian optimization reaffirmed the central question of whether CNNs can infer temporally ancient and nearly neutral evolutionary scenarios, confirming that CNNs can reliably detect these selection signals with high accuracy across most scenarios. Specifically, the model achieved accuracies ranging from 69.10% to 88.88%. However, it struggled with recent and weak selection, where accuracy was limited to 50.40%, a particularly challenging case to infer even with optimization. Importantly, selection signals that were previously difficult to detect were now elucidated, providing opportunities for further research. Compared to manual tuning, Bayesian optimization offers a more efficient method for improving accuracy, achieving better results more quickly and effectively.

We also experimented with increasing the number of optimisation trials from the original 30 to 50. While this yielded some improvements in certain scenarios, it also exacerbated overfitting in others, particularly in cases where overfitting had already been observed. Notably, the performance on recent and weak selection remained poor despite the additional trials, suggesting that a more expansive hyperparameter optimization space might be necessary for future work (see Appendix A)

While we know that Bayesian optimisation outperforms other hyperparameter optimization methods, the potential of transfer learning raises an important question: could transfer learning lead to quicker and better hyperparameter optimization with a fraction of the computational complexity? This possibility is explored briefly in the appendix (see Appendix B).

7.2 Human Pigmentation Genes

This segment of the experiment confirmed the validity and efficacy of the CNNs in detecting signatures of selection. It was consistently observed that the predictions from testing the CNNs on known variants under selection matched the selection coefficients and selective pressures reported in the literature. Additionally, due to their accuracy in matching the strength of selection, it is reasonable to conclude that the inference for the timing of selection can also be considered fairly accurate. Furthermore, the observation that genes related to human pigmentation were generally selected around 50,000 Kya, as the results suggest aligns logically with the second larger migrations of AMH out of Africa [68]. During these migrations, ancient humans encountered colder climates, where lighter pigmentation would have been favored for better adaptation. As humans migrated farther from the equator, they encountered lower levels of ultraviolet radiation (UVR). In these regions, lighter skin, which allows more UVR to penetrate and produce vitamin D, would have been advantageous for survival, as people with darker skin require more UVR to produce sufficient vitamin D [134].

7.3 Autophagy Dysregulation Genes

The findings indicate that majority of autophagy dysregulation genes have display moderate selective pressure around 50,000 years ago, a pattern consistently observed across these variants. This timing aligns with the literature, as it coincides with the second major migration out of Africa [68]. This migration likely increased the effective population size (Ne), thereby enhancing the efficacy of natural selection in fixing beneficial mutations and eliminating deleterious ones [139], especially when exposed to novel environmental pressures. These factors likely intensified the selection for autophagy-related variants, facilitating adaptation to the new conditions. The selective pressure may have been driven by the need for enhanced survival mechanisms in these unfamiliar environments, despite potential long-term consequences, such as increased susceptibility to diseases that manifest later in life.

One possible explanation for why these disease-associated genes were under positive selection during this period is the concept of antagonistic pleiotropy [135]. This hypothesis suggests that these variants may have conferred protective advantages early in life, which were crucial for survival and reproduction in the new environments. However, this same biological compromise could have led to detrimental effects later in life, as seen in conditions like hypertension, breast cancer, and ankylosing spondylitis, which typically manifest post-reproductive age [136][137][138]. The early-life advantages likely outweighed the later-life costs, driving the positive selection of these variants during this critical period of human evolution.

7.4 Limitations and steps for future improvement

One limitation of the study was the prevalence of overfitting, which persisted despite efforts to mitigate it through on-the-fly simulations. Repeated use of the same data batches across multiple epochs, while beneficial for training, resulted in overfitting, even with high validation accuracy. Addressing this challenge will require generating larger datasets to increase the number of data batches. While the results are promising, there are several areas for further enhancement of the experimental design. For instance, expanding the hyperparameter space and increasing the number of trials could boost the model's predictive power. Additionally, training the models on other demographic models, such as African or Asian populations, would enhance generalizability.

Moreover, the project did not incorporate multiclass or regression classification methods, which restricted our ability to quantify the timing of selection events more precisely. Instead, our approach could only infer these events without providing precise temporal resolution. Another limitation could be that we didn’t extend ImaGene framework to test for introgression between populations as the time periods of the datasets reflect periods of history when Homo Sapiens existed with other hominids. Testing for potential gene introgression from these hominids could have provided critical insights, potentially elucidating whether the selected genes originated from differing populations or even distinct hominid species.

Furthermore, while the study's methodology is robust, there is an underlying concern, as also noted by Flagel et al. (2019) [58], regarding the extent to which human-designed preprocessing methods might inadvertently shape what machine learning models perceive. This prompts us to reconsider whether our models are truly learning from the data or merely reflecting our biases. Therefore, it would be prudent to explore distinct approaches that remove reliance on preprocessing steps, such as developing order-agnostic deep learning neural networks. Such networks could employ operations that are invariant to the arrangement of haplotypes, ensuring that the models are unaffected by the sequence in which the haplotypes are provided.

The utility of permutation-invariant networks is well-supported in the literature, as demonstrated by Chan et al. (2018) [64], who used a network invariant to row permutations to detect recombination hotspots, thereby improving both accuracy and training speed in the CNN. Although this method was not applied to positive selection inference, it shows promising signs for broader applications. Moreover, the current ImaGene pipeline, which works in tandem with the coalescent simulator msms [88], could be significantly enhanced by integrating advanced genealogical data representations into forward-time genomic simulations, as suggested by Torada et al. (2019) [13]. Also as previously mentioned the study could have had increased scope experimenting with different pre-trained models as this could have led to increase accuracy and better HPO's faster.

8.0 Conclusion

By extending the capabilities of ImaGene, we have demonstrated its effectiveness in detecting and inferring signatures of natural selection across various temporal scales, including both recent and ancient evolutionary scenarios. This study validates the efficacy of CNNs for detecting ancient selection scenarios and nearly neutral moderate selection in recent periods, which have been proven difficult to deduce using traditional methods and even some modern techniques like deep learning. Moreover, this work overcomes the temporal myopia often seen in the literature, where models are typically trained only to infer selection for recent and strong selective sweeps. This accomplishment represents a significant milestone, as it has resulted in the development of a classifier capable of detecting the presence of selection. With our eight models spanning 10-100kya, we can now semi-infer the timing and strength of selection for any given genomic region of interest.

Future work will focus on refining model accuracy, incorporating diverse demographic models, and exploring new biological questions, to deduce if other genomic regions are under directional selective pressure. Additionally, integrating state-of-the-art data representations—such as those capturing genealogical histories in forward-time simulations—could further improve computational efficiency and generalizability, solidifying ImaGene as a powerful tool for evolutionary inference. Furthermore, future efforts should explore multiclass or regression classification approaches, enabling a more precise, quantitative determination of the timing of selection rather than relying on proxy inference.

9.0 Bibliography

[1] Eline DL, Matteo F, Bo L, Kelley H, Zijun X, Zhou et al. "Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears 2014. https://doi.org/https://doi.org/10.1016/j.cell.2014.03.054.

[2] Ilardo M, Nielsen R. Human adaptation to extreme environmental conditions. Current Opinion in Genetics & Development 2018;77–82. https://doi.org/10.1016/j.gde.2018.07.003.

[3] Trucchi EABMLAISVLNEB et al. "Ancient genomes reveal early A farmers selected common beans while preserving diversity, Benazzo A, Lari M, Lob A, Vai S, Nanni L, et al. 2021. https://doi.org/10.1038/s41477-021-00848-7.

[4] Orna M, Leslie H, Adam RB, Amit I, Carolin K, Carlos DB, et al. Natural selection on genes that underlie human disease susceptibility 2008. https://doi.org/https://doi.org/10.1016/j.cub.2008.04.074.

[5] Nielsen Rasmus "Molecular signatures of natural selection. " AnnuRevG. Molecular signatures of natural selection 2005. https://doi.org/https://doi.org/10.1146/annurev.genet.39.073003.112420.

[6] Voight BF, Kudaravalli S, Wen X, Pritchard JK. A Map of Recent Positive Selection in the Human Genome. PLoS Biology 2006. https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0040072 (accessed August 22, 2024).

[7] Tom R, Benjamin CJ, Peter DK. Detecting positive selection in the genome 2017. https://doi.org/https://doi.org/10.1186/s12915-017-0434-y.

[8] Nikolaos A, Pavlos P. Detecting positive selection in populations using genetic data. Humana, New York: Statistical population genomics; 2020. https://doi.org/https://doi.org/10.1007/978-1-0716-0199-0_5.

[9] Sirén J, Kaski S. Local dimension reduction of summary statistics for likelihood-free inference - Statistics and Computing. Statistics and Computing 2019;559–570. https://doi.org/10.1007/s11222-019-09905-w.

[10] Christian P, Jean-Marie C, Jean-Michel M, Natesh SPillai "Lack of confidence in approximate B computation model choice. Lack of confidence in approximate Bayesian computation model choice 2011. https://doi.org/https://doi.org/10.1073/pnas.1102900108.

[11] Jiaping W, Michael K, Nhung H, Hyong HL, Iain M, Sara Mathieson "Automatic inference of demographic parameters using generative adversarial networks 2021. https://doi.org/10.1111/1755-0998.13386.

[12] Bárbara D, Débora YB, Diogo M, Aida MAndrés "Inferring balancing selection from genome-scale data. Inferring balancing selection from genome-scale data 2023. https://doi.org/https://doi.org/10.1093/gbe/evad032.

[13] Torada L, Lorenzon L, Beddis A, Isildak U, Pattini L, Mathieson S, et al. ImaGene: a convolutional neural network to quantify natural selection from genomic data - BMC Bioinformatics. BMC Bioinformatics 2019;20. https://doi.org/10.1186/s12859-019-2927-x.

[14] Benjamin M, Emilia H-S, Rasmus N. Distinguishing between Selective Sweeps from Standing Variation and from a De Novo Mutation 2012. https://doi.org/https://doi.org/10.1371/journal.pgen.1003011.

[15] Aaron J, Peter RW, Rasmus Nielsen "An approximate full-likelihood method for inferring selection and allele frequency trajectories from D sequence data 2019. https://doi.org/10.1371/journal.pgen.1008384.

[16] Wolfgang Stephan "Maximum-likelihood methods for detecting recent positive selection and localizing the selected site in the genome, Li H 2005. https://doi.org/doi: 10.1534/genetics.105.041368.

[17] Reid Nancy "Approximate likelihoods. Approximate likelihoods n.d.

[18] Manolo F, Isabel AB, Monique R, Fernando FF, Nigel PT, Daniela CZ, et al. Coalescent-based species delimitation meets deep learning: insights from a highly fragmented cactus system. Molecular Ecology Resources 2022. https://doi.org/https://doi.org/10.1111/1755-0998.13534.

[19] Michael GB, Oscar EG, Olivier François "Approximate B computation (ABC) in practice. Approximate Bayesian computation (ABC) in practice 2010. https://doi.org/https://doi.org/10.1016/j.tree.2010.04.001.

[20] Johann B, Gilles Louppe "The frontier of simulation-based inference. The frontier of simulation-based inference 2020. https://doi.org/https://doi.org/10.1073/pnas.1912789117.

[21] Prangle D. Summary Statistics in Approximate Bayesian Computation. arXivOrg 2015. https://doi.org/https://doi.org/10.48550/arXiv.1512.05633.

[22] Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 1989;585–595. https://doi.org/10.1093/genetics/123.3.585.

[23] Fu Y-Xin "Statistical properties of segregating sites. Statistical properties of segregating sites 1995. https://doi.org/https://doi.org/10.1006/tpbi.1995.1025.

[24] Justin C, Chung-I. Wu "Hitchhiking under positive D selection. Hitchhiking under positive Darwinian selection 2000. https://doi.org/https://doi.org/10.1093/genetics/155.3.1405.

[25] laume. "Frequency spectrum neutrality tests: one for all and all for one. Frequency spectrum neutrality tests: one for all and all for one 2009. https://doi.org/https://doi.org/10.1534/genetics.109.104042.

[26] Dilber E, Terhorst J. Robust detection of natural selection using a probabilistic model of tree imbalance. Genetics 2022;220. https://doi.org/10.1093/genetics/iyac009.

[27] Gabor T, Eva C, Janos M, Stephen TSherry "The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations 2004.

[28] Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 1995;783–796. https://doi.org/10.1093/genetics/140.2.783.

[29] Bachtrog D. Evidence that positive selection drives Y-chromosome degeneration in Drosophila miranda - Nature Genetics. Nature Genetics 2004;518–522. https://doi.org/10.1038/ng1347.

[30] Abigail W, Xianyun M, Rui M, Tom B, Megan JW, Colleen GJ, et al. Identifying positive selection candidate loci for high-altitude adaptation in Andean populations 2009.

[31] Marc B, Dalila P, Xianyun M, Joshua MA, Rui M, hen WS et al. "Identifying signatures of natural selection in T and A populations using dense genome scan data 2010.

[32] Kelly JK. A Test of Neutrality Based on Interlocus Associations. Genetics 1997;1197–1206. https://doi.org/10.1093/genetics/146.3.1197.

[33] Rasmus Nielsen "Linkage disequilibrium as a signature of selective sweeps. Linkage disequilibrium as a signature of selective sweeps 2004.

[34] Jessica L, Yu-Ping P, Shivani M, Jeffrey DJ. The impact of equilibrium assumptions on tests of selection 2013.

[35] Sarah E, Reuben JP, Timothy JS, Andrew Collins "Sequencing era methods for identifying signatures of selection in the genome. Sequencing era methods for identifying signatures of selection in the genome 2019.

[36] Manjit P, Anuradha P, Divya R, Sonali SN, K. AS, Kaiho K, et al. Machine-learning prospects for detecting selection signatures using population genomics data 2022.

[37] Zachary D, Skylar YL, Faraz F, Roy HC, Chengxiang Z, Miles JE, et al. big data: astronomical or genomical? 2015.

[38] Shawn E, Richard MMyers "Advancements in next-generation sequencing. Advancements in next-generation sequencing 2016.

[39] Paul Marjoram "Approximately sufficient statistics and B computation. Approximately sufficient statistics and Bayesian computation 2008.

[40] Christoph L, Laurent Excoffier "Efficient approximate B computation coupled with M chain MC without likelihood. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood 2009.

[41] Jones N. Nature: Computer science: The learning machines. (2014) 2014.

[42] Jeffrey DJ, Wolfgang Stephan "Searching for footprints of positive selection in whole-genome S data from nonequilibrium populations 2010.

[43] Scott W, Yuseob K, Melissa JH, Andrew GC, Carlos Bustamante "Genomic scans for selective sweeps using S data. Genomic scans for selective sweeps using SNP data 2005.

[44] Alexandros S, Pavlos Pavlidis "OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets 2012.

[45] Nitin U, Eran H, Vineet Bafna "Learning natural selection from the site frequency spectrum. Learning natural selection from the site frequency spectrum 2013.

[46] Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, et al. Signals of recent positive selection in a worldwide sample of human populations. Genome Research 2009;826–837. https://doi.org/10.1101/gr.087577.108.

[47] Corinna C, Lawrence DJ, Yann L, Vladimir V. Boosting and other machine learning algorithms n.d.

[48] Lin K, Yang Z, Li H. Illustrations for evolBoosting. Https://WwwPicbAcCn/Evolgen/Softwares/Download/evolBoosting/evolBoosting%20manual_100Pdf 2012. https://www.picb.ac.cn/evolgen/softwares/download/evolBoosting/evolBoosting%20manual_1.0.0.pdf (accessed August 23, 2024).

[49] Lin K, Li H, Schlötterer C, Futschik A. Distinguishing Positive Selection from Neutral Evolution: Boosting the Performance of Summary Statistics. Genetics 2011;229–244. https://doi.org/10.1534/genetics.110.122614.

[50] Kevin R, Jeffrey DJensen "Controlling the false-positive rate in multilocus genome scans for selection. Controlling the false-positive rate in multilocus genome scans for selection 2007.

[51] Bühlmann P, Hothorn T. Boosting Algorithms: Regularization, Prediction and Model Fitting. Statistical Science 2007;22. https://doi.org/10.1214/07-sts242.

[52] Pierre L, Giovanni MD, Manu U, Hafid L, Jaume B, Johannes Engelken "Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations 2015.

[53] Daniel R, Andrew DKern "S/HIC: robust identification of soft and hard sweeps using machine learning. S/HIC: robust identification of soft and hard sweeps using machine learning 2016.

[54] Oscar EG, Matteo Fumagalli "Deep learning in population genetics. Deep learning in population genetics 2023.

[55] Daniel R, Andrew DKern "Supervised machine learning for population genetics: a new paradigm. Supervised machine learning for population genetics: a new paradigm 2018.

[56] Yun SSong "Deep learning for population genetic inference. Deep learning for population genetic inference 2016.

[57] Andrew D, Daniel RSchrider "diploS/HIC: an updated approach to classifying selective sweeps 2018.

[58] Yaniv B, Daniel RSchrider "The unreasonable effectiveness of convolutional neural networks in population genetic inference 2019.

[59] Alessandro S, Matteo Fumagalli "Distinguishing between recent balancing selection and incomplete sweep using deep neural networks 2021.

[60] Cecil RM, Sugden LA. On convolutional neural networks for selection inference: Revealing the effect of preprocessing on model learning and the capacity to discover novel patterns. PLoS Computational Biology 2023; e1010979. https://doi.org/10.1371/journal.pcbi.1010979.

[61] Szandała T. Unlocking the black box of CNNs: Visualising the decision-making process with PRISM 2023.

[62] Nandita R, Philipp WM, Erkan OB, Dmitri APetrov "Recent selective sweeps in NAD melanogaster show signatures of soft sweeps. Recent Selective Sweeps in North American Drosophila Melanogaster Show Signatures of Soft Sweeps 2015.

[63] Nguembang F, Fabrizio R, Giorgio B, Emiliano T. Identification of natural selection in genomic data with deep convolutional neural network 2021.

[64]A Likelihood-Free Inference Framework for Population ... PubMed 2018. https://pubmed.ncbi.nlm.nih.gov/33244210/ (accessed August 23, 2024).

[65] Gower G, Picazo PI, Fumagalli M, Racimo F. Detecting adaptive introgression in human evolution using ... eLife 2021;10. https://doi.org/10.7554/elife.64669.

[66] M. E, Kasper M, David Enard "Versatile detection of diverse selective sweeps with flex-sweep. Versatile detection of diverse selective sweeps with flex-sweep 2023.

[67] Marta M, Robert Foley "Multiple dispersals and modern human origins. Multiple dispersals and modern human origins 1994.

[68] Towards a theory of modern human origins: geography, demography, and diversity in recent human evolution - PubMed. PubMed 1998. https://doi.org/10.1002/(sici)1096-8644(1998)107:27.

[69] Sankararaman S, Patterson N, Li H, Pääbo S, Reich D. The Date of Interbreeding between Neandertals and Modern Humans. PLoS Genetics 2012; e1002947. https://doi.org/10.1371/journal.pgen.1002947.

[70] Swapan M, Nick P, David Reich "The combined landscape of D and N ancestry in present-day humans 2016.

[71] Carey John "Unearthing the origins of agriculture. Unearthing the origins of agriculture 2023.

[72] Allison AC "Protection afforded by sickle-cell trait against subtertian malarial infection. Protection afforded by sickle-cell trait against subtertian malarial infection 1954.

[73] S. U, M. J, F. S, M. P, D. PKwiatkowski "A comparison of case-control and family-based association methods: T example of sickle-cell and malaria. A comparison of case-control and family-based association methods: The example of sickle-cell and malaria 2005.

[74] nic P "How malaria has affected the human genome and what human genetics can teach us about malaria. How malaria has affected the human genome and what human genetics can teach us about malaria 2005.

[75] Huttley GA, Easteal S, Southey MC, Tesoriero A, Giles GG, McCredie MR, et al. Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees - Nature Genetics. Nature Genetics 2000;410–413. https://doi.org/10.1038/78092.

[76] Melissa A, John DP, Christina JR, Gary KO, Elaine AOstrander "Understanding missense mutations in the B gene: an evolutionary approach 2003.

[77] Thomas GB. BRCA1 and BRCA2 in breast cancer 2001.

[78] Dezheng H, Catherine S, Barbara N, Anselm H, M. CL, Suh-Yuh W, et al. A signature of balancing selection in the region upstream to the human UGT2B4 gene and implications for breast cancer risk 2011.

[79] Linda E, Gareth TJ, Alan GM, Christina D, Roger DS, Gary JMacfarlane "Global prevalence of ankylosing spondylitis. Global prevalence of ankylosing spondylitis 2014.

[80] Álvaro FF, Guillermo Mariño "Pathogenic single nucleotide polymorphisms on autophagy-related genes. Pathogenic single nucleotide polymorphisms on autophagy-related genes 2020.

[81] Hypertension 2023. https://www.who.int/news-room/fact-sheets/detail/hypertension#:~:text=The%20number%20of%20adults%20with,risk%20factors%20in%20those%20populations. (accessed August 23, 2024).

[82] Facts and figures | Breast Cancer UK. Breast Cancer UK 2024. https://www.breastcanceruk.org.uk/about-breast-cancer/facts-figures-and-qas/facts-and-

figures/#:~:text=Global%20breast%20cancer%20statistics&text=In%202022%2C%20there%20were%20around,deaths%20among%20women%20(1). (accessed August 23, 2024).

[83] Patient information sheet sIgAD. Https://MftNhsUk/App/Uploads/Sites/7/2018/04/SelectiveIgAdeficiencypatientsheetPdf n.d. https://mft.nhs.uk/app/uploads/sites/7/2018/04/SelectiveIgAdeficiencypatientsheet.pdf (accessed August 23, 2024).

[84] How to Know if Your Machine Learning Model Has Good Performance n.d. https://www.obviously.ai/post/machine-learning-model-performance (accessed August 23, 2024).

[85] Lior Shamir "Evaluation of the benchmark datasets for testing the efficacy of deep convolutional neural networks 2021.

[86] Team K. Keras: Deep Learning for humans n.d. https://keras.io/ (accessed August 23, 2024).

[87] Paul B, Jianmin C, Zhifeng C, Andy D, Jeffrey D, hieu D et al. "{TensorFlow}: a system for {Large-S machine learning. " I 12th U symposium on operating systems design and implementation (OSDI 16). {TensorFlow}: a system for {Large-Scale} machine learning n.d.

[88] Joachim Hermisson "MSMS: a coalescent simulation program including recombination, graphic structure and selection at a single locus 2010.

[89] Kingman JFC. The coalescent. Stochastic Processes and Their Applications 1982;235–248. https://doi.org/10.1016/0304-4149(82)90011-4.

[90] msms User Manual n.d. https://www.mabs.at/fileadmin/user_upload/p_mabs/Manual.pdf (accessed August 23, 2024).

[91] Elisabetta C, Donata Luiselli "Inferring signatures of positive selection in whole-genome sequencing data: an overview of haplotype-based methods.2022.

[92] John Haigh "The hitch-hiking effect of a favourable gene. The hitch-hiking effect of a favourable gene 2007.

[93] Ellen ML, Yongtao G, Matthew S, Graham C, Molly Przeworski "Variation in human recombination rates and its genetic determinants 2011.

[94] Richard Durbin "Revising the human mutation rate: implications for understanding human evolution 2012.

[95] Alexandre M, Nandita RG, Michael DeGiorgio "Detection and classification of hard and soft sweeps from unphased genotypes by multilocus genotype identity 2018.

[96] Van der W, Johannes LS, Juan N-I, François B, Joshua DW, Neil Y, et al. scikit-image: image processing in Python 2014.

[97] Christopher K. Understanding the Benefits of Image Augmentations 2023.

[98] Aja H, Chris JM, Arthur G, Laurent S, George VDD, an S et al. "Mastering the game of G with deep neural networks and tree search 2016.

[99] GPU machine types n.d. https://cloud.google.com/compute/docs/gpus (accessed August 23, 2024).

[100] Bellman RE. Adaptive Control Processes. Princeton University Press; 2015.

[101] Yoshua Bengio "Random search for hyper-parameter optimization. Random search for hyper-parameter optimization. 2012.

[102] Michael D, Richard JB, William JConover "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code 2000.

[103] Ilya A, V. MSaleev "An economic method of computing Lp. An economic method of computing LPτ-sequences 1979.

[104] Bayesian optimization is superior to random search for machine learning hyperparameter tuning Analysis of the black-box optimization challenge 2020 n.d.

[105] Hans-Paul S. Evolution strategies–a comprehensive introduction 2002.

[106] Beyer H-G, Sendhoff B. Evolution Strategies for Robust Optimization. IEEE Conference Publication | IEEE Xplore 2006. https://doi.org/10.1109/CEC.2006.1688465.

[107] Donald R, Matthias S, William JW. Efficient global optimization of expensive black-box functions 1998.

[108] Holger HH, Kevin L-B. Sequential model-based optimization for general algorithm configuration n.d.

[109] Eduardo C, Daniel H-L. Dealing with categorical and integer-valued variables in bayesian optimization with gaussian processes 2020.

[110] Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press; 2016.

[111] Shotaro S, Toshihiko Y, Takeru O, Masanori Koyama "Optuna: A next generation hyperparameter optimization framework. Optuna: A next generation hyperparameter optimization framework n.d.

[112] Adesh B, Asif S. A comparative study of hyper-parameter optimization tools n.d.

[113] Andrew Y "Feature selection, L 1 vs. L 2 regularization, rotational invariance. " IP of the twenty-first international conference on M learning. Feature selection, L1 vs. L2 regularization, and rotational invariance n.d.

[114] Xiangyu Z, Shaoqing R, Jian Sun "Deep residual learning for image recognition. " IP of the I conference on computer vision and pattern recognition. Deep residual learning for image recognition n.d.

[115] Dropout: a simple way to prevent neural networks from overfitting: The Journal of Machine Learning Research: Vol 15, No 1. The Journal of Machine Learning Research 2014. https://doi.org/10.5555/2627435.2670313.

[116] i. “Convolutional neural networks.” In Deep learning for robot perception and cognition. Convolutional neural networks n.d.

[117] Richard A, David LD. Human pigmentation genes under environmental selection 2012.

[118] Pardis CS, Nick P, Trisha V, Steve FS, Jared AD, Matthew R, et al. Genetic signatures of strong recent positive selection at the lactase gene 2004.

[119] Mandrekar JN "Receiver operating characteristic curve in diagnostic test assessment. Receiver operating characteristic curve in diagnostic test assessment 2010.

[120] Concepción MA, Ángel Gil "Genetics of lactose intolerance: an updated review and online interactive world maps of phenotype and genotype frequencies 2020.

[121] Toshiyuki H, Kosuke MTeshima "Power of neutrality tests for detecting natural selection. Power of neutrality tests for detecting natural selection 2023.

[122] Pleuni SPennings "Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Soft sweeps: molecular population genetics of adaptation from standing genetic variation 2005.

[123] Jong-Wha Jung. "Small molecule inhibitors for Unc-51-like autophagy-activating kinase targeting autophagy in cancer." 2023.

[124] S. F. Schaffner, P. Sabeti. "Evolutionary adaptation in the human lineage." 2008.

[125] Parsa M., Schuman C.D., Date P., Rose D.C., et al. "Hyperparameter optimization in binary communication networks for neuromorphic deployment." 2020.

[126] Kevin RT, Jaime A, Peter LRalph. "Efficient Pedigree Recording for Fast Population Genetics Simulation." 2018.

[127] Benjamin C., Jared G., Jerome K., Philipp W. M., Peter L., & Ralph P. (2019). Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. **Genetics, 211**(3), 647-661.

[128] Saeed S, Dmitry BG, Peter RM, Lawrence OH. Challenges for the repeatability of deep learning models. 2020

[129] Patel CDBUSRPSPKMNC et al. "DBGC: D generic convolution block for object recognition. DBGC: Dimension-based generic convolution block for object recognition 2022.

[130] Mizuho, N., Richard, K. G. D., & Kaori, T. (2018). Convolutional neural networks: An overview and application in radiology.

[131] Prajit R, Zoph B, Le QV. Searching for activation functions. 2017.

[132] Limin W, Yufei Z, Xuming H, Muhammet D, Parmar M. A review of convolutional neural networks in computer vision. 2024.

[133] Ryan N. An introduction to convolutional neural networks. 2015.

[134] António MS, Brian M, Isabel A, Cláudia M, Emily C, Mark DS, et al. The timing of pigmentation lightening in Europeans. 2013.

[135] **Stephen, C., Randolph, M. N., Diddahally, R. G., & Ellison, P. T.** (2010). Evolutionary perspectives on health and medicine. Evolutionary Perspectives on Health and Medicine.

[136] High Blood Pressure a Common Condition as We Age. *n.d.* https://memorialregionalhealth.com/health-topics/cardiology/high-blood-pressure-common-condition-age/#:~:text=While%20only%2025%25%20of%20men,surpass%20men%20after%20age%2075.

[137] Key Statistics for Breast Cancer. *n.d.* https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html#:~:text=Breast%20cancer%20mainly%20occurs%20in,cancer%20are%20younger%20than%2045.

[138] Ankylosing Spondylitis. *n.d.* https://www.hopkinsmedicine.org/health/conditions-and-diseases/ankylosing-spondylitis.

[139] Vahdati AR, Sprouffske K, Wagner A. Effect of Population Size and Mutation Rate on the Evolution of RNA Sequences on an Adaptive Landscape Determined by RNA Folding. *International Journal of Biological Sciences*. 2017;13(9):1138–1151. https://doi.org/10.7150/ijbs.19436.

10.0 Appendix A

List of abbreviations:

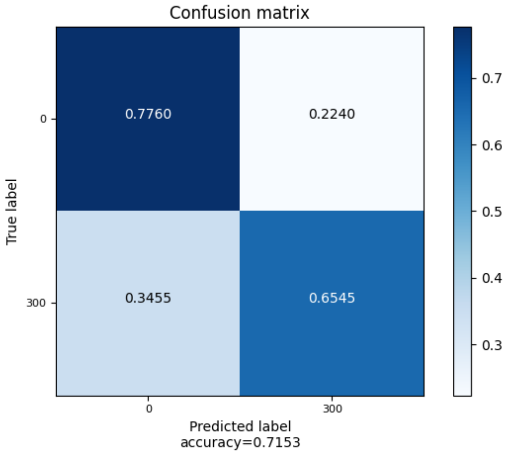
- ABC: Approximate Bayesian Computation
- AUC: Area Under the Curve
- ATG13_S: ATG13 gene related to Selective IgA Deficiency
- ATG13_B: ATG13 gene related to Breast Cancer
- CNN: Convolutional Neural Network
- DL: Deep Learning
    - FCNN: Fully Connected Neural Network
    - ML: Machine Learning
    - msms: Coalescent Simulation Program
    - Ne: Effective Population Size
    - ReLU: Rectified Linear Unit
    - ROC: Receiver Operating Characteristic
    - SFS: Site Frequency Spectrum
    - SNP: Single Nucleotide Polymorphism
    - LD: Linkage Disequilibrium
    - FLOPS: Floating Point Operations
    - AMH : Anatomically Modern Humans

Table. Table showing the parameters and floating-point operations for the architectures

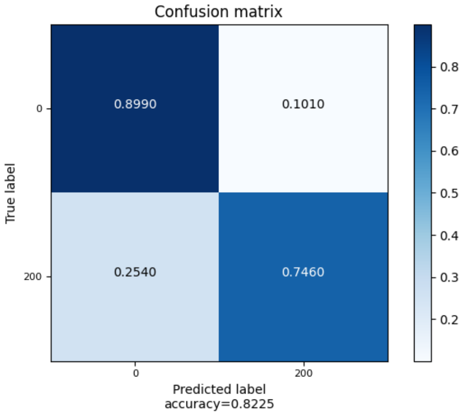
Computational Metrics

	Parameters	Floating-Point operations
Baseline	4,173,473	2.72125313×10^{8}
Recent and moderate	490,931	3.17954059×10^{8}
Recent and weak	3,528,564	2.43104488×10^{9}
Intermediate strong	3,351,216	3.63706201×10^{8}
Intermediate moderate	5,092,744	6.67099285×10^{8}
Intermediate weak	203,971	$2.581355128 \times 10^{9}$
Ancient strong	3,091,595	7.25121082×10^{8}
Ancient moderate	2,492,909	5.36664571×10^{8}
Ancient weak	16,409,518	8.84400196×10^{8}

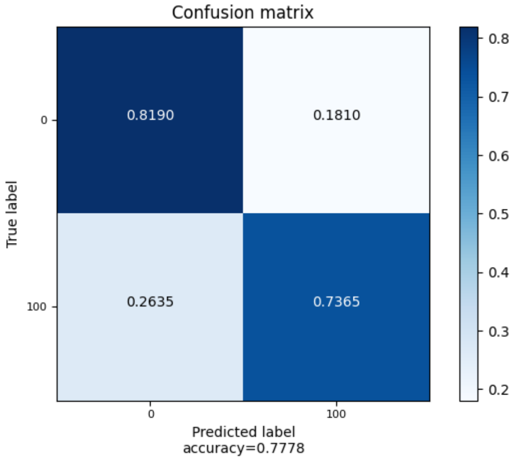
Note. This table presents the parameters and floating-point operations (FLOPs) for CNN models optimized using Bayesian optimization, where the hyperparameters were tailored to specific evolutionary scenarios representing distinct datasets. These models were optimized (trained) according to the characteristics of each scenario. Additionally, the table includes the baseline low-complexity CNN model's parameters and FLOPs for comparison. For scenarios effectively addressed by the baseline, such as "Recent and Weak," re-evaluation was not performed. FLOPs and parameter metrics were calculated using standard TensorFlow methods to assess the computational complexity, scalability, and efficiency of each model



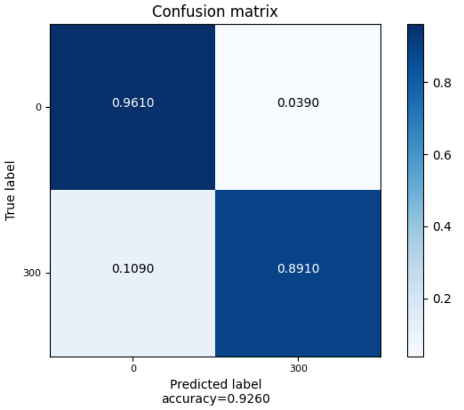
Confusion matrix for ancient and strong selection with Bayesian opmtisation with 50 trials.



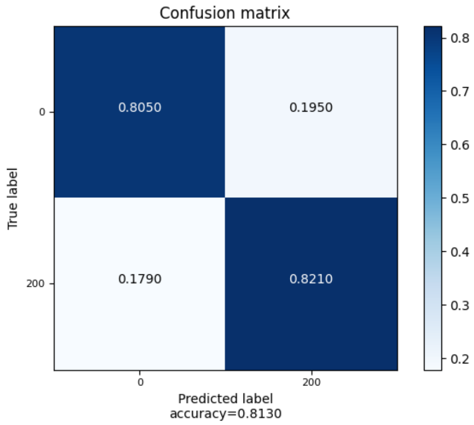
Confusion matrix for ancient and moderate selection with Bayesian opmtisation with 50 trials.



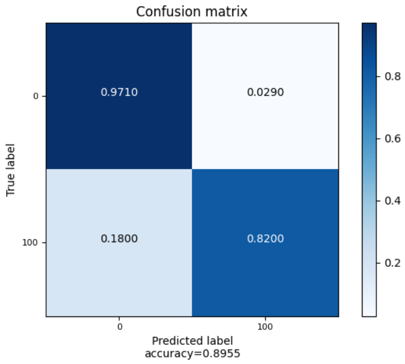
Confusion matrix for ancient and weak selection with Bayesian opmtisation with 50 trials.



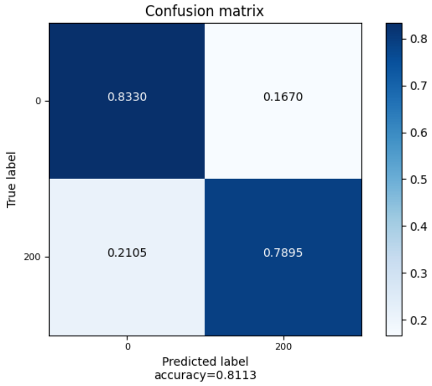
Confusion matrix for intermediate and strong selection with Bayesian opmtisation with 50 trials.



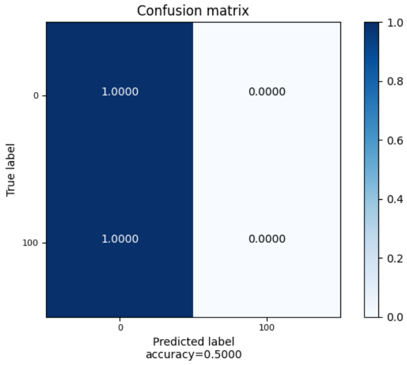
Confusion matrix for intermediate and moderate selection with Bayesian opmtisation with 50 trials.



Confusion matrix for intermediate and weak selection with Bayesian opmtisation with 50 trials.

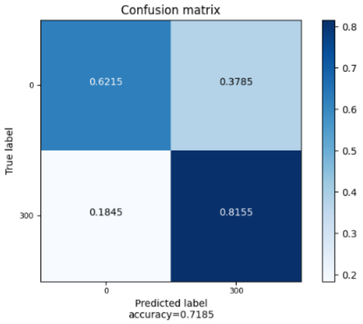


Confusion matrix for recent and moderate selection with Bayesian opmtisation with 50 trials.

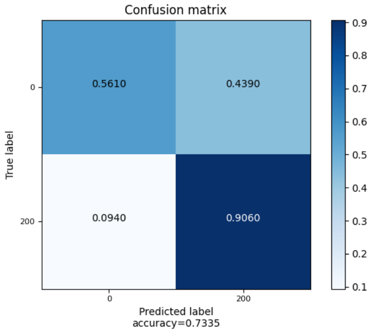


Confusion matrix for recent and weak selection with Bayesian opmtisation with 50 trials.

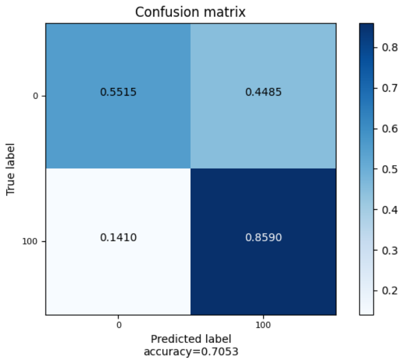
11.0 Appendix B



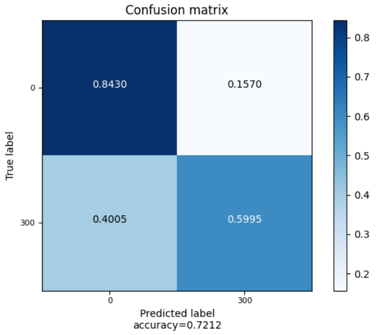
Confusion matrix for ancient and strong selection using pre-trained VGG-16 network with ImageNet weights.



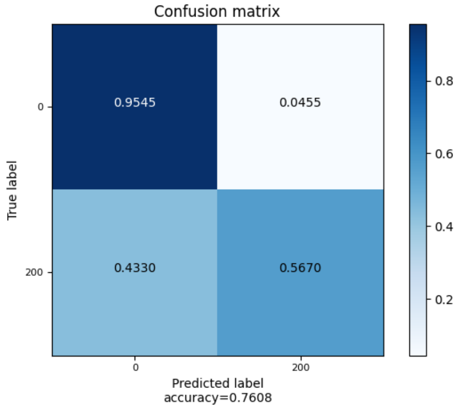
Confusion matrix for ancient and moderate selection using pre-trained VGG-16 network with ImageNet weights.



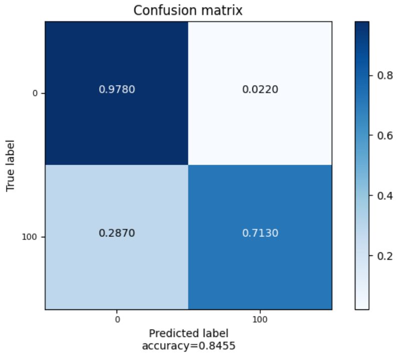
Confusion matrix for ancient and weak selection using pre-trained VGG-16 network with ImageNet weights.



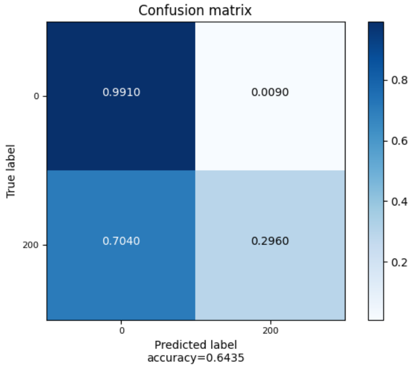
Confusion matrix for intermediate and strong selection using pre-trained VGG-16 network with ImageNet weights.



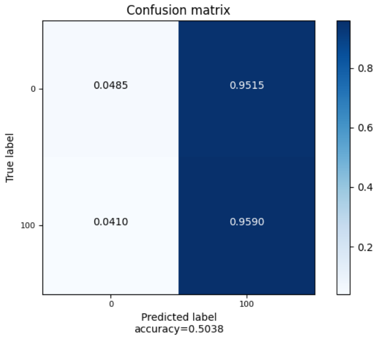
Confusion matrix for intermediate and moderate selection using pre-trained VGG-16 network with ImageNet weights.



Confusion matrix for intermediate and weak selection using pre-trained VGG-16 network with ImageNet weights.



Confusion matrix for recent and moderate selection using pre-trained VGG-16 network with ImageNet weights.



Confusion matrix for recent and weak selection using pre-trained VGG-16 network with ImageNet weights.