Evaluating Convolutional Neural Networks and Transformer Architectures for Image-Based Prediction of Protein Localization in Eukaryotic Cells

---

Faculty of Health Sciences

University Of Cape Town

Department of Computational Biology

MSc. Med (by dissertation) in Bioinformatics

Sibongiseni Letticia Msipa

MSPSIB002

SUBMITTED TO THE UNIVERSITY OF CAPE TOWN

In fulfilment of the requirements for this degree

supervisor: Dr. Sinkala Musalula – musalula.sinkala@uct.ac.za, Department of Integrative Biomedical Sciences, University of Cape Town

2025

Declaration

I, *Sibongiseni Msipa*, hereby declare that the work on which this dissertation/thesis is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signed by candidate

Signature: _____

Date: 10 January 2025

Acknowledgements

I am deeply grateful to my parents for their unwavering support, encouragement, and the countless opportunities they have provided me. Their belief in my dreams has been instrumental in my journey.

I am also thankful to the University of Cape Town (UCT) and its generous donors for the partial scholarship during my first year of Masters. Additionally, I am indebted to The UCT Scholarship for the scholarship in my final year, which significantly eased the financial burden of my education. Their commitment to advancing education has enabled me to pursue my academic aspirations.

Special appreciation goes to Dr. Musalula Sinkala, whose steadfast support provided me with the motivation to overcome challenges and remain focused on my goals. Your belief in my potential has been a constant inspiration.

I am also grateful to my Family and Palesa Tjale, who have contributed to my academic and personal growth throughout this journey. Your insights and guidance have been invaluable.

Lastly, I acknowledge the countless individuals who, in various ways, have contributed to my academic and personal development. Your contributions have been deeply appreciated and have played a significant role in shaping this thesis. This work is a culmination of the collective support and encouragement I have received, and I am sincerely thankful to each and every person who has been part of this remarkable journey.

List of Abbreviations

3D-CNN - Three-Dimensional Convolutional Neural Network

Adam - Adaptive Moment Estimation (optimizer)

AI - Artificial Intelligence

AUC: Area Under the Curve

AutoML - Automated Machine Learning

avg – Average

CNN - Convolutional Neural Network

CNN - Convolutional Neural Network

DAPI - 4',6-Diamidino-2-Phenylindole (a fluorescent stain for DNA)

DenseNet121 - Densely Connected Convolutional Network with 121 layers

DL - Deep Learning

DL: Deep Learning

DNN - Deep Neural Network

DTI - Drug-Target Interaction

ER - Endoplasmic Reticulum

GAN - Generative Adversarial Network

GNN - Graph Neural Network

GPU - Graphics Processing Unit

Grad-CAM - Gradient-weighted Class Activation Mapping

Grad-CAM: Gradient-weighted Class Activation Mapping

HPA - Human Protein Atlas

Hugging Face Company providing transformer-based models and tools

IF - Immunofluorescence

IHC - Immunohistochemistry

ImageNet - Large Visual Database for Image Recognition

InceptionV3 - Inception architecture version 3

Keras - High-level neural networks API, often used with TensorFlow

KNN - K-Nearest Neighbor

LASSO - Least Absolute Shrinkage and Selection Operator

LBP - Local Binary Pattern

LQP - Local Quinary Pattern

LSTM - Long Short-Term Memory

LTP - Local Ternary Pattern

Matplotlib - Python plotting library

ML - Machine Learning

MLM - Masked Language Modeling

MLP - Multilayer Perceptron

MSA - Multiple Sequence Alignment

NLP - Natural Language Processing

OpenCV - Open Source Computer Vision Library

PPI - Protein-Protein Interaction

PTM - Post-Translational Modification

RF - Random Forest

RGB - Red, Green, Blue (color channels)

RNA-seq - RNA Sequencing

RNN - Recurrent Neural Network

ROC: Receiver Operating Characteristic

SIFT - Scale-Invariant Feature Transform

SLF - Subcellular Location Features

SURF - Speeded-Up Robust Features

SVM - Support Vector Machine

Swin Transformer - Shifted Window Transformer

Swin-Tiny - Swin Transformer Tiny Model

Swiss-Prot - Swiss Protein Knowledgebase

TensorFlow - Open-source machine learning framework

TrEMBL - Translated European Molecular Biology Laboratory

TSV - Tab-Separated Values

UniProt - Universal Protein Resource

ViT - Vision Transformer

ViT-Base - Vision Transformer Base Model

Xception - Extreme Inception

List of Figures

LIST OF TABLES

Abstract

Background: Accurate prediction of protein subcellular localization is critical for understanding protein function and guiding experimental research. Recent advances in deep learning have enabled high-throughput image-based methods to tackle this problem by leveraging large-scale immunofluorescence microscopy datasets. The aim of this study is to comparatively evaluate convolutional neural network (CNN) architectures and Transformer-based models for the multi-label classification of protein subcellular localization in eukaryotic cells, using large-scale immunofluorescence image datasets.

Methods: In this study, we comparatively evaluated convolutional neural network (CNN) architectures (DenseNet121, Xception, and InceptionV3) and transformer-based models (Vision Transformer and Swin Transformer) for multi-label classification of protein localization in eukaryotic cells. Using 12,565 immunofluorescence images from the Human Protein Atlas—representing 15 subcellular compartments—we performed transfer learning by replacing the final layers of pretrained ImageNet models to accommodate multi-label output. All models were trained with iterative stratification to handle class imbalance and evaluated on held-out test images.

Results and discussion: Our findings indicate that CNN-based models, particularly DenseNet121 and Xception, achieve the highest overall accuracy and F1-scores, successfully recognizing both abundant and underrepresented classes. In contrast, transformers demonstrated variable performance. While the Swin Transformer surpassed the Vision Transformer, neither consistently matched CNN performance—likely reflecting the data requirements and hyperparameter sensitivity of transformer architectures. Visualization techniques (Grad-CAM in CNNs and attention maps in transformers) confirmed that well-performing models localize salient features to biologically relevant regions, suggesting they learn meaningful morphological cues

Conclusion: These results underscore CNNs' suitability for subcellular localization analysis with moderate-scale datasets, while transformers may require more extensive tuning or larger training sets to reach comparable accuracy. Our findings suggest that CNNs, especially DenseNet121 and Xception, exhibit superior performance over transformer models in predicting protein localization. CNN-based models demonstrate higher accuracy and interpretability, positioning them as preferred choices for advancing functional proteomics and computational drug discovery.

Keywords: Protein Subcellular Localization, Convolutional Neural Networks, Vision Transformers, Deep Learning, Multi-Label Classification.

Chapter One

1 Introduction

The subcellular localization of a protein refers to its specific position within a cell. This information is critical in fields such as molecular cell biology, proteomics, and systems biology, as it sheds light on the protein's function. The location of a protein is essential for its role biologically because the different cellular compartments offer unique chemical environments, potential interaction substrates or partners, necessary for diverse functions. Consequently, understanding a protein's subcellular localization is a key step in deciphering its function (Wang & Wei, 2022). Many cellular processes, such as the transportation of nucleocytosolic transcription factors (Stewart, 2007), the localization of mitochondrial proteins during apoptosis (Mayor & Pagano, 2007), and the uptake of endocytic cell-surface cargo receptors, depend on precise protein localization (Garapati et al., 2020; Wang & Wei, 2022). Misplaced proteins can disrupt their functionality, potentially leading to diseases such as cancers (Wang & Wei, 2022), neurodegenerative disorders (Barmada et al., 2010; Ziff et al., 2023), and metabolic conditions (Xiang et al., 2006; Lundberg et al., 2019). Thus, identifying protein subcellular locations can aid in anticancer therapies (Lomenick et al., 2011Therefore, determining the subcellular locations of proteins can benefit anticancer treatment (Lomenick et al., 2011) and improve target identification for drug discovery (Wang & Wei, 2022).

Moreover, protein localization provides insights into the cellular processes of hypothetical and newly discovered proteins (Tahir et al., 2014). For example, drug molecules can more easily access plasma membrane proteins and secreted proteins due to their surface location (Tscherepanow et al., 2008). Subcellular localization also plays a role in the early diagnosis of diseases, as abnormal localization is often observed in conditions like Alzheimer's and cancer. Additionally, accurately identifying protein localization helps determine the functional environment of proteins (Chen et al., 2006). In summary, precise knowledge of protein localization is crucial for drug identification and efficacy (Tahir et al., 2014)..

Several experimental techniques are available for analyzing protein localization. Quantitative mass spectrometry enables the identification of proteins across different fractions (Jakobsen et al., 2011;

Christoforou et al., 2016; Itzhak et al., 2016; Orre et al., 2019). Temporal and spatially resolved proteomic maps can be obtained in living cells using targetable peroxidase (Rhee et al., 2013; Hung et al., 2014; Lee et al., 2016). In addition, methods such as immunofluorescence and high-resolution confocal microscopy have facilitated the visual estimation of protein localization within individual cells (Chong et al., 2015; Barbe et al., 2008; Stadler et al., 2010; Thul et al., 2017; Burns et al., 2017). These experimental methods yield high-resolution location of targeted proteins for researchers, enabling direct observation to uncover biological processes and metabolic mechanisms (Xiao et al., 2024).

However, these wet lab experimental methods also have several significant drawbacks: they often require expensive equipment and time-consuming steps, making them costly for large-scale studies (Xiao et al., 2024). These problems are exacerbated because newly discovered proteins are increasing exponentially in the post-genomic era.

To address these issues, computational methods and machine learning (ML) techniques have become essential for predicting protein subcellular localization. Researchers have developed numerous bioinformatics-based prediction systems integrated with ML approaches to localize a wide range of proteins (Chebira et al., 2007; Hamilton et al., 2007; Khan et al., 2008; Khan et al., 2011; Lin et al., 2007; Murphy et al., 2003; Nanni & Lumini, 2008; Nanni et al., 2010a; Zhang et al., 2009).

Machine Learning (ML), a subfield of Artificial Intelligence (AI), has driven much of the recent progress in AI applications. ML focuses on using data to solve specific tasks—such as predicting protein properties based on known data from other proteins by leveraging algorithms trained on data sets to create self-learning models that are capable of predicting outcomes and classifying information without human intervention (Xiao et al., 2024; Coursera Staff, 2024). A significant subset of ML, known as deep learning (DL), has particularly revolutionized predictive capabilities, enabling models to achieve unprecedented accuracy across a variety of domains. Deep learning is an advanced branch of machine learning that excels at handling a broader spectrum of data types—including text and unstructured forms like images—while needing even less human oversight (McKinsey & Company, 2024). This often leads to results that surpass the accuracy of traditional machine learning. The technique relies on neural networks that mimic the behavior of human brain neurons, processing information through several layers (McKinsey & Company, 2024). As data

moves through these successive layers, the network is capable of identifying progressively more intricate patterns and features (McKinsey & Company, 2024).

In protein localization prediction, DL methods have been instrumental, setting new standards in fields such as object detection (He et al., 2015), semantic segmentation (Girshick et al., 2014), image captioning (Vinyals et al., 2015), and biological applications spanning regulatory genomics (Alipanahi et al., 2015; Zhou & Troyanskaya 2015; Kelley et al., 2016) to electron microscopy (Ciresxan et al., 2012, 2013; Tan et al. 2015; Angermueller et al. 2016; Rampasek & Goldenberg 2016). In many tasks, these models even outperform human performance, underscoring their transformative potential (Jiang et al., 2021).

Given these advancements, determining protein subcellular localizations using computational methods integrated with ML techniques have become essential. In this context, researchers have worked to develop various bioinformatics-based prediction systems, using ML approaches, to localize a wide range of proteins (Chebira et al., 2007; Hamilton et al., 2007; Khan et al., 2008; Khan et al., 2011; Lin et al., 2007; Murphy et al., 2003; Nanni and Lumini, 2008; Nanni et al., 2010a; Zhang et al., 2009). Importantly, experimental and computational methods for protein localization are complementary. Experimental annotations often serve as ground truth labels for computational approaches, and computational models are trained on these data to predict the localization of other proteins (Jiang et al., 2021). Given their cost-effectiveness, automation, and high-throughput capabilities, computational methods are invaluable for large-scale protein subcellular localization characterization (Jiang et al., 2021).

This rapidly evolving field of deep learning has driven numerous successful real-world applications, including image recognition (LeCun et al., 1998), gaming (Silver et al., 2017), and autonomous vehicles (Bojarski et al., 2016). In recent years, Convolutional Neural Networks (CNNs) have become the dominant method for image classification, segmentation, and object detection, particularly in data-intensive areas like biological image analysis. Unlike traditional techniques like Loc-CAT, which depend on manually crafted features, CNNs directly process raw images and learn hierarchical feature representations in an end-to-end manner. These representations can be further refined through transfer learning (Jiang et al, 2021). Transfer learning in machine learning involves taking the knowledge a neural network has acquired from one task and applying it as a starting point for training the model on a different task (Linkon et al.,

2021; Ahmed et al., 2023). This approach leverages an existing model trained on a large dataset (such as ImageNet) to serve as a feature extractor within a convolutional neural network. The method involves removing the final fully connected (classifier) layer of the pre-trained network and repurposing the remaining layers for a new task (Linkon et al., 2021). Instead of retraining the entire network, only a new classifier is trained on top of the extracted features, which considerably accelerates the training process (Linkon et al., 2021; Ahmed et al., 2023). This capability allows CNNs to efficiently and scalably capture cellular localization patterns (Liimatainen et al., 2021; Ehteshami et al., 2017; Krizhevsky et al., 2012; Ouyang et al., 2019).

CNNs have been the predominant architecture in biological image analysis, but the recently introduced Vision Transformers (ViTs) (Dosovitskiy et al., 2020) have yet to be fully explored in this field. Unlike CNNs, which rely on convolutional operations, ViTs are free from convolution-induced biases, allowing the model to learn global features and capture complex relationships in the data. Given the success of transformers in natural language processing (Devlin et al., 2018; Liu et al., 2019), this raises the question of whether ViTs could be the next breakthrough in image recognition and potentially replace CNNs in the future (Gai et al., 2024).

Current CNN-based approaches have achieved high accuracy but often lack interpretability, while transformer-based methods remain underexplored in the context of protein localization (Dosovitskiy et al., 2020). Furthermore, existing research has not fully evaluated the comparative strengths and limitations of these architectures in the same experimental setting (Gai et al., 2024).. This study addresses that gap by systematically comparing CNN and transformer-based models, aiming to provide insights into how each method can be optimized for protein localization prediction.

While CNNs have demonstrated outstanding performance in many image-based tasks, transformers offer a global attention mechanism that captures complex relationships and long-range dependencies (Casola, Lauriola & Lavelli, 2022; Chen, 2022; Pratap, 2023). Their success in natural language processing suggests they could be equally transformative for image-based applications, including protein subcellular localization (Dosovitskiy et al., 2020). By directly comparing CNNs with Vision Transformers, this work provides a clearer understanding of the conditions under which transformer-based models might outperform or complement CNN

approaches, thereby guiding future research toward more interpretable and scalable solutions (Gai et al., 2024).

Here, I aim to compare CNNs and Transformer models for predicting protein localization in eukaryotic cells through image analysis, specifically using immunofluorescence techniques. While CNNs have proven effective in biological image analysis, their ability to capture local patterns through convolutional layers has some limitations, especially in learning global features and complex relationships. In contrast, ViTs utilize self-attention mechanisms to capture both local and global information, offering a potentially more robust approach to protein localization prediction (Pratap, 2023). By directly processing raw images without relying on manually crafted features, CNNs and ViTs can autonomously learn the intricate patterns of protein localization in a variety of subcellular compartments. This research will explore the effectiveness of both architectures, comparing their ability to accurately predict protein localization in eukaryotic cells and potentially identify new trends in applying deep learning to bioinformatics. Ultimately, my work seeks to improve our understanding of protein localization through computational methods, paving the way for more efficient and scalable approaches to bioinformatics in cell biology.

Chapter Two- Literature Review

2 Protein Localization and Machine Learning-Based Prediction Models

The subcellular localization of a protein refers to its specific position within a cell. This information is crucial in molecular cell biology, proteomics, and systems biology, as it provides insights into the protein's function. The location where a protein resides is essential for its roles biologically, because the different cellular compartments offer unique chemical environments, such as redox conditions and pH balance, and potential interaction partners, necessary for diverse functions. Understanding a protein's subcellular localization is therefore a critical step in deciphering its function (Wang & Wei, 2022). Many cellular processes, including transport of nucleocytosolic transcription factors (Stewart, 2007), the localization of mitochondrial proteins during apoptosis (Mayor & Pagano, 2007), and the uptake of endocytic of cell-surface cargo receptors, depend on precise protein localization (Garapati et al., 2020; Wang & Wei, 2022). Misplaced proteins can disrupt their functionality, potentially leading to diseases such as cancers (Wang & Wei, 2022), neurodegenerative disorders (Barmada et al., 2010; Ziff et al., 2023), and metabolic conditions (Xiang et al., 2006; Lundberg et al., 2019). Consequently, identifying protein subcellular locations can aid in anticancer therapies (Lomenick et al., 2011) and improve target identification for drug development (Wang & Wei, 2022).

Additionally, protein localization helps describe various cellular processes involving hypothetical and newly discovered proteins (Tahir et al., 2014). For instance, drug molecules can more easily access plasma membrane proteins and secreted proteins due to their surface location (Tscherepanow et al., 2008). Subcellular localization also plays a role in the early diagnosis of diseases, as abnormal localization is often observed in conditions like Alzheimer's and cancer (Chen et al., 2006). Furthermore, accurately identifying protein localization helps determine the functional environment of proteins (Chen et al., 2006). In summary, precise knowledge of protein localization is vital for drug identification and efficacy (Tahir et al., 2014).

Several experimental techniques are available for analyzing protein localization. Quantitative mass spectrometry enables the identification of proteins across different cellular fractions (Jakobsen et al., 2011; Christoforou et al., 2016; Itzhak et al., 2016; Orre et al., 2019). Temporal and spatially resolved proteomic maps can be generated in living cells using targetable peroxidase (Rhee et al., 2013; Hung et al., 2014; Lee et al., 2016). Techniques such as immunofluorescence and high-resolution confocal microscopy allow for the visualization of protein localization within individual cells (Chong et al., 2015; Barbe et al., 2008; Stadler et al., 2010; Thul et al., 2017; Burns et al., 2017; Kohnhorst et al., 2016; Feng et al., 2017). Immunoelectron microscopy, which uses labeled antibodies against target proteins, is considered a gold standard for providing high-resolution electron microscopy (Xiao et al., 2024). These methods offer high-resolution data, enabling researchers to directly observe and uncover biological processes and metabolic mechanisms (Xiao et al., 2024). However, these experimental approaches have significant drawbacks: they are time-consuming and require expensive equipment, which makes them impractical for large-scale studies (Xiao et al., 2024). In this post-genomic era this challenge is exacerbated by the exponential growth of newly discovered proteins (Xiao et al., 2024).

Take the UniProt Database as an example - over the past decade, the disparity between reviewed and unreviewed proteins has notably widened (Figure 1A). Specifically, as shown in Figure 1B, the latest UniProt release (2024_01) reveals that a substantial majority of entries are unreviewed proteins within TrEMBL. Given this scale, relying solely on wet lab experiments to determine subcellular localization across vast datasets from multiple species (Figure 1C) is increasingly impractical. However, the extensive repository of accurately annotated protein data available in databases (Figure 1D) provides valuable resources for developing robust predictive models (Xiao et al., 2024). Notably, the smaller size of Swiss-Prot compared to TrEMBL reflects the intensive manual curation applied to Swiss-Prot entries.

In contrast, TrEMBL includes computationally analyzed records, resulting in a significant volume of protein sequences pending annotation before integration into Swiss-Prot (Xiao et al., 2024). Fortunately, manual curation demands might be reduced by validating transcript-translated sequences through proteomics. An example of this approach is seen in the Human Protein Atlas (HPA), where RNA sequencing (RNA-seq data) was used to support immunofluorescence

subcellular localization results. In this context, computational models, especially Artificial Intelligence (AI) driven techniques adept at managing large-scale datasets, can provide substantial advantages (Xiao et al., 2024).

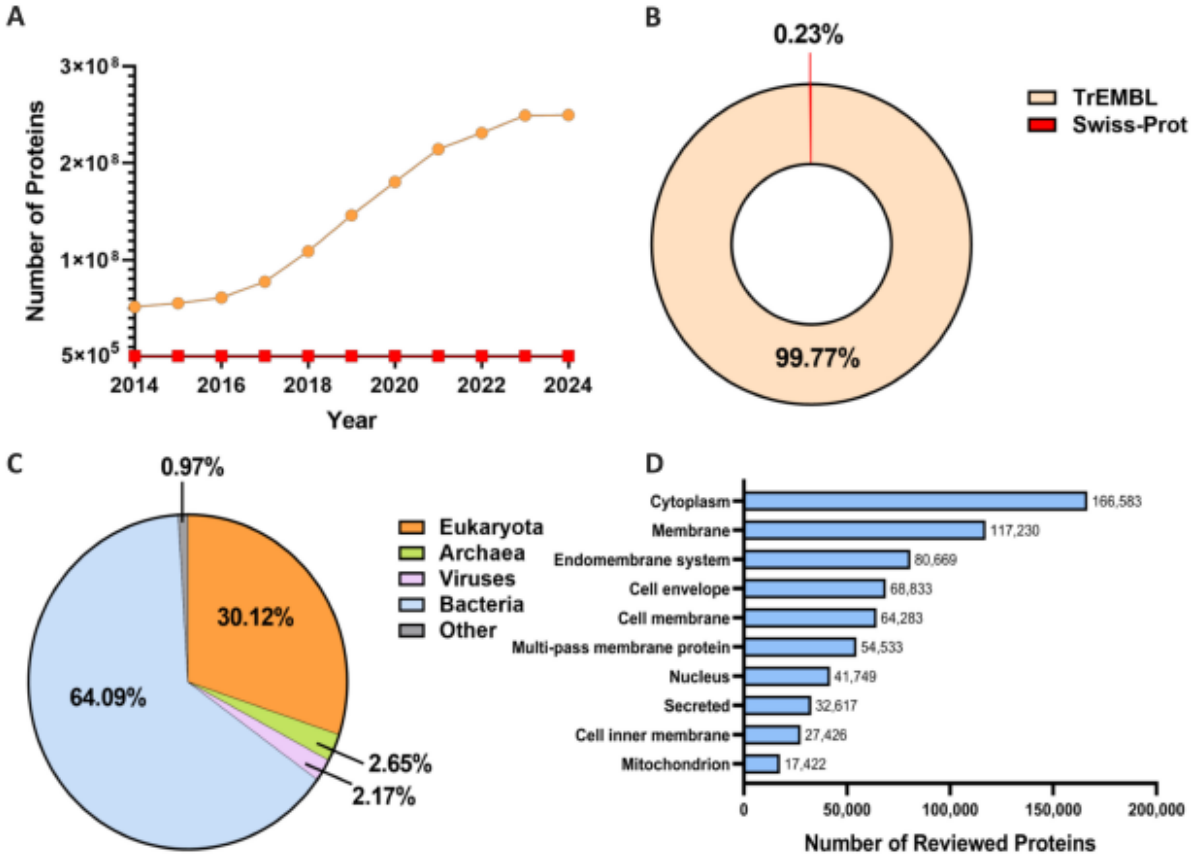


Figure 1: ***Growth trends of proteins in UniProt databases over time****.*

A) Depicts how the number of proteins has increased over time in TrEMBL (unreviewed proteins) versus Swiss-Prot (reviewed proteins). Newly identified uncharacterized proteins grow at a much faster rate than those experimentally verified. (B) Shows the ratio of newly added proteins in each database for the 2024_01. version. (C) Highlights the taxonomic distribution of the protein sequences. (D) Illustrates the number of proteins found in the top 10 subcellular locations.

In this context, computational methods coupled with machine learning (ML) techniques are essential for determining protein subcellular localization. ML, a subfield of AI, has driven much of the recent progress in AI applications. It uses data to solve specific tasks, such as predicting protein properties based on known data from other proteins. A significant subset of ML, known as deep learning (DL), has revolutionized predictive capabilities, enabling models to achieve unprecedented accuracy across various domains. In protein localization prediction, DL methods have set new benchmarks in fields such as object detection (He et al., 2015), semantic segmentation

(Girshick et al., 2014), image captioning (Vinyals et al., 2015), and biological applications ranging from regulatory genomics (Alipanahi et al., 2015; Zhou & Troyanskaya, 2015; Kelley et al., 2016) to electron microscopy (Ciresan et al., 2012, 2013; Tan et al., 2015; Angermueller et al., 2016; Rampasek & Goldenberg, 2016). These models often outperform human performance, highlighting their transformative potential. Researchers have developed numerous bioinformatics-based prediction systems integrated with ML approaches to localize a wide range of proteins (Chebira et al., 2007; Hamilton et al., 2007; Khan et al., 2008; Khan et al., 2011; Lin et al., 2007; Murphy et al., 2003; Nanni & Lumini, 2008; Nanni et al., 2010a; Zhang et al., 2009).

In recent decades, in-silico techniques for predicting protein subcellular localization have advanced rapidly. These techniques can be broadly categorized into three approaches: (i) knowledge-based techniques, which draw on annotations of proteins from various databases; (ii) sequence-based techniques, that rely solely on amino acid sequences; and (iii) image-based methods, which leverage bioimages to extract features indicative of subcellular localization. The substantial progress in ML and DL, along with the growing availability of experimentally determined localization data, imaging records in public databases and functional annotations, has made these computational frameworks increasingly accurate and efficient. This review focuses exclusively on the third category—image-based methods—highlighting the potential of ML-driven techniques to enhance protein subcellular localization predictions based on bioimage data.

2.1 Bioimage-Based Methods

Imaging data shows direct visual evidence of protein localization within different cell components, allowing precise and accurate location determination. Through imaging processing, computational models can analyze the spatial distribution of proteins at the single-cell level and quantify their localization patterns (Xiao et al., 2024). Florescence-based imaging techniques are the standard in protein localization investigations because of the ability to visualize intracellular proteins, either through the expression of fluorescent fusion proteins or the recognition of target proteins by fluorophore-detected, antibody-based techniques (Xiao et al., 2024). The complexity of images offers different levels of features, which also require multiple preprocessing steps, deep classification models, and a longer running time to deal with for better performance (Liimatainen et al., 2021; Xiao et al., 2024)

The task of protein localization through bioimages is a form of image annotation, where the assigned labels correspond to different cellular compartments (Long et al., 2019). Unlike conventional image annotation, this problem is more complex since labels are associated with proteins instead of individual images. A single protein may appear in multiple images taken from different experimental conditions, requiring careful consideration of localization patterns present across the entire dataset (Long et al., 2019). Additionally, the number of images per protein varies, leading to datasets of different sizes. This challenge aligns with a specialized machine learning approach known as multi-instance learning (Foulds & Frank, 2010; Zhou, 2004).Furthermore, proteins can localize to multiple cellular compartments, making this a multi-instance multi-label problem (Zhou et al., 2012). In multi-instance learning, each data point consists of multiple instances grouped together in a "bag," and its label depends on the overall characteristics of these instances. A sample is classified as positive if at least one instance in the bag is positive; otherwise, it is considered negative (Long et al., 2019). Meanwhile, in multi-label learning, a single sample can be linked to multiple labels instead of just one (Long et al., 2019).

2.2 Bioimage-Based Features

Representing proteins with 2D images, rather than amino acid sequences, offers a more intuitive and concise approach for determining subcellular localization. As microscopic imaging technology has rapidly advanced, bioimage-based methods for protein localization have gained significant attention (Wang & Wei, 2022; Xiao et al., 2024). Improved computational hardware, particularly advances in graphics processing units (GPUs), now enables researchers to tackle complex calculations required for such analysis (Kobayashi et al., 2022). This progress, alongside developments in neural network architectures, has greatly accelerated DL applications in bioimage analysis (Jiang et al., 2021).

To support high-quality bioimaging data, the HPA program was launched in 2003 to map all human proteins across cells, tissues, and organs (The Human Protein Atlas, n.d; Thul et al., 2017). This open-access database contains imaging data, mass spectrometry-based proteomics, and transcriptomics, among others (HPA, n.d). The subcellular section of HPA, which has recently been updated to version 23, contains detailed information on protein expressions and spatial distributions for 13,147 genes, making it a vital resource for developing computational methods in

bioimaging (Thul & Lindskog, 2018; Ouyang et al., 2019). This study will draw from the HPA data to build and train ML models for predicting the subcellular localization of proteins. Immunofluorescence (IF) and immunohistochemistry (IHC) images are widely used in recent studies as benchmark sources for training and testing bioimage models (Xiao et al., 2024).

Predicting protein subcellular localization involves identifying specific cellular compartments—such as the cytoplasm, nucleus, or vesicle—where a protein resides, achieved through image analysis and classification (Xiao et al., 2024). By examining distinct visual characteristics in protein images, bioimage-based models aim to categorize proteins into established subcellular patterns (Jiang et al., 2021). This process of subcellular localization prediction can be framed as a multiclass classification problem involving two core steps: feature extraction and classification (Jiang et al., 2021). Effective feature extraction is essential, as it significantly impacts the ML model's performance. Traditionally, image descriptors for protein localization prediction have been categorized into global and local features (Galar et al., 2011; Wen et al., 2015).

Subcellular location features (SLF) are divided into global and local categories (Xu et al., 2018). Global features, which include DNA distribution information and overall image texture, provide a broad description of spatial structures within the image. Examples include morphological features, local binary patterns (LBP) (Nanni et al., 2012), and Haralick and Zernike moments (Tahir et al., 2012). Zernike moments (Tahir et al., 2012). Zernike moments (Tahir et al., 2012) provide features that remain unchanged under image transformations like translation and rotation. These features are often integrated with others, such as Haralick texture descriptors, to enhance the prediction of protein subcellular localization (Tahir et al., 2012). Additionally, DNA distribution characteristics play a role in evaluating the spatial organization of human cells, contributing to improved variability analysis in research (Newberg & Murphy, 2008; Xu et al., 2013; Yang et al., 2014b; Zuo et al., 2020). The LBP technique, which relies on predefined spatial patterns, is used to generate histograms that help characterize the structural aspects of protein localization (Nanni et al., 2012; Tahir et al., 2013; Yang et al., 2014a). Various LBP extensions, such as Local Quinary Pattern (LQP) and Local Ternary Pattern (LTP) , further refine the extraction of local features useful for predicting protein localization (Tahir et al., 2012; Yang et al., 2014b). Moreover, the Haralick texture descriptor is widely applied in pattern analysis, providing statistical insights through metrics like correlation, contrast, and entropy derived from the gray-level co-occurrence

matrix (Haralick et al., 1973). In contrast, local features capture finer micro-patterns that global features may overlook. Techniques such as invariant LBP, Speeded-Up Robust Features (SURF) and the bag-of-visual-words approach, are commonly applied for this purpose (Glorot et al., 2011; Shao et al., 2017). Additionally, the scale-invariant feature transform (SIFT) is useful for identifying salient points in fluorescence images, enhancing performance when combined with global features (Ouyang et al., 2019). This combined approach ensures robust analysis of fluorescence bioimages, facilitating accurate protein localization predictions (Ouyang et al., 2019).

However, these manually designed features represent only basic, low-level image information and are constrained by current IF imaging knowledge, limiting localization performance advancements. In contrast, with its robust feature representation and learning capabilities, DL has recently been widely adopted for predicting protein subcellular locations.

2.3 Bioimage-Based AI Methods for Protein Localization

Image-related methods can be roughly organized into three phases based on the algorithms and the number of data types used: conventional ML, DL, and complex fusion methods. Figure 2 shows the development of these models from simple to complicated.

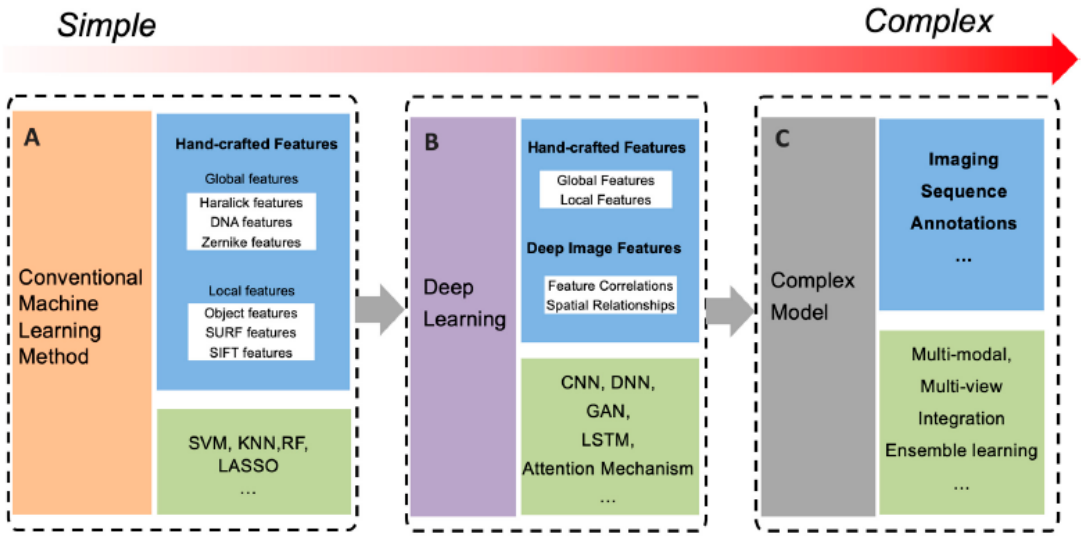


*Figure 2: **Categories of computational approaches for protein localization***

There are three main categories of computational methods for processing imaging data, represented here by a red arrow denoting the increasing complexity of predictive models and signaling advancements toward more sophisticated frameworks. Features used for model training are highlighted in a blue rectangle, and location prediction algorithms are noted in green. First, Conventional Machine Learning Methods rely on manually crafted features that capture both global and local aspects of images, then employ simpler predictive models. Next, Deep Learning Methods add deep image representations derived from neural networks, combining them with hand-crafted features. Finally, Complex Fusion Models incorporate multiple data types—such as sequence information, annotation text, and image data—into a single model, offering a more comprehensive and interpretable approach to protein subcellular localization. Commonly used techniques and acronyms include SURF (Speeded Up Robust Features), SIFT (Scale-Invariant Feature Transform), SVM (Support Vector Machine), KNN (K-Nearest Neighbor), RF (Random Forest), LASSO (Least Absolute Shrinkage and Selection Operator), CNN (Convolutional Neural Network), DNN (Deep Neural Network), GAN (Generative Adversarial Network), and LSTM (Long Short-Term Memory). Deep neural network implementation typically serves as the starting point, providing a foundation for more advanced or specialized enhancements

Traditional ML models for protein subcellular localization have historically relied on hand-crafted features for classification (Liu et al., 2020; Newberg & Murphy, 2008; Zou et al., 2023). These methods manually select and engineer specific image features based on domain knowledge. For instance, Li et al. (2012) utilized structured latent variables in a logistic regression model to capture complex patterns in various image regions, while Ullah et al. (2021) employed a two-layer feature selection model with an SVM using both radial basis function and linear kernels. While these approaches can achieve high accuracy, they are often sensitive to noise and variations in imaging data, resulting in decreased model robustness. Spatial relationships embedded in images are also rarely detected due to manual feature engineering, which limits scalability (Xiao et al., 2024).

In contrast, DL models have revolutionized image-based protein localization by automatically extracting features from raw data without requiring manual feature engineering (Chandra et al., 2023). DL models are distinct in that they operate as end-to-end systems, removing the need for manual feature engineering typically required in traditional ML (Chandra et al., 2023). Instead of relying on predefined input features, deep neural networks autonomously learn to interpret raw protein image data and convert it into meaningful internal representations. This approach enables highly accurate protein subcellular localization predictions by embedding data processing and classification within a single step, enhancing the model's precision and adaptability to complex image data (Khan et al., 2019; Xiao et al., 2024). This transition from hand-crafted to learned features has significantly improved model performance, especially on large, complex datasets, as it allows for deeper and more nuanced feature extraction (Newberg & Murphy, 2008).

During image preprocessing, deep neural networks (DNNs) utilize processed image segmentation as input for multi-layer convolutional neural networks (ML-CNN) to enhance feature extraction (Pärnamaa & Parts, 2017). Some models incorporate both low- and high-level bioimage features to conduct a more comprehensive analysis. For multi-label classification, conventional CNNs are modified using a specialized learning approach that takes into account both label-attribute relationships and label-label dependencies. This technique helps refine predictions by accurately determining the final subcellular localization (Wang & Wei, 2022; Su et al., 2021).

More recently, Transformer-based models have shown promising results in bioimage analysis by utilizing attention mechanisms to capture global image contexts (Vaswani et al., 2017; Chandra et al., 2023). Unlike CNNs, which rely on local receptive fields, Transformers employ self-attention to create a combined representation of the entire image. This capability has been demonstrated by Long et al. (2020), who integrated self-attention and multi-head attention mechanisms to enhance the feature representations in immunohistochemistry images. Vision Transformers (ViTs), as implemented by Wang and Wei (2022), are particularly suited for multi-scale feature extraction, as they enable the model to learn and aggregate information across various scales, which is beneficial for capturing complex spatial patterns within cells (Dosovitskiy et al., 2021). Although transformer models have yet to be widely applied to protein subcellular localization, studies such as Zhao et al. (2024) suggest that these models, when optimized with techniques like graph and resolution-based transformers, have great potential to fully exploit the depth of information within imaging data.

Despite their strengths, transformer models still face limitations in efficiency and computational demands compared to CNNs, particularly in large-scale analyses. This review will compare the efficacy of CNNs and transformer-based models for predicting protein localization in eukaryotic cells through bioimage analysis.

In the following section, we delve into how DL techniques specifically address protein subcellular localization prediction challenges.

2.4 Deep Learning for Protein Subcellular Localization Prediction

2.4.1 Introducing Neural Networks in Protein Localization

The architecture of a neural network—its layout, depth, and connectivity patterns—defines its learning capabilities and inherent assumptions about the data it processes. Different architectures, including multilayer perceptrons (MLPs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs), have been developed to address diverse data types and specific applications (Koumakis, 2020). Each architecture brings unique strengths to deep learning applications, especially when tailored to specific prediction tasks.

Among these, MLPs, CNNs, and RNNs (Figure 3) are some of the most widely used and effective architectures, often applied individually or in combination to solve complex tasks. For example, CNNs are widely used for feature extraction in protein fold recognition, while RNNs are valuable for sequence-based data, such as protein sequences (Koumakis, 2020). Additionally, word embedding models like Word2Vec are employed in natural language processing tasks to generate meaningful embeddings of words or amino acid sequences, often using the continuous bag-of-words and skip-gram models (Liu et al., 2020).

2.4.2 MLP, CNN, and RNN Architectures in Deep Learning

MLPs are straightforward neural networks consisting of an input layer, several hidden layers, and an output layer (Chandra et al., 2023). They are versatile and can handle different types of inputs but require dense connectivity, where each neuron in one layer is connected to every neuron in the next. CNNs, on the other hand, are optimized for image data, employing convolution operations to automatically extract hierarchical features from input images (Mikolov et al., 2011; Mikolov et al., 2013a). These convolution layers apply learned filters to detect local features, while pooling layers reduce the dimensionality, preserving essential patterns. While CNNs excel in image and audio analysis, they also offer promising results in analyzing sequence data, including protein sequences and DNA-protein binding (Zeng et al., 2016).

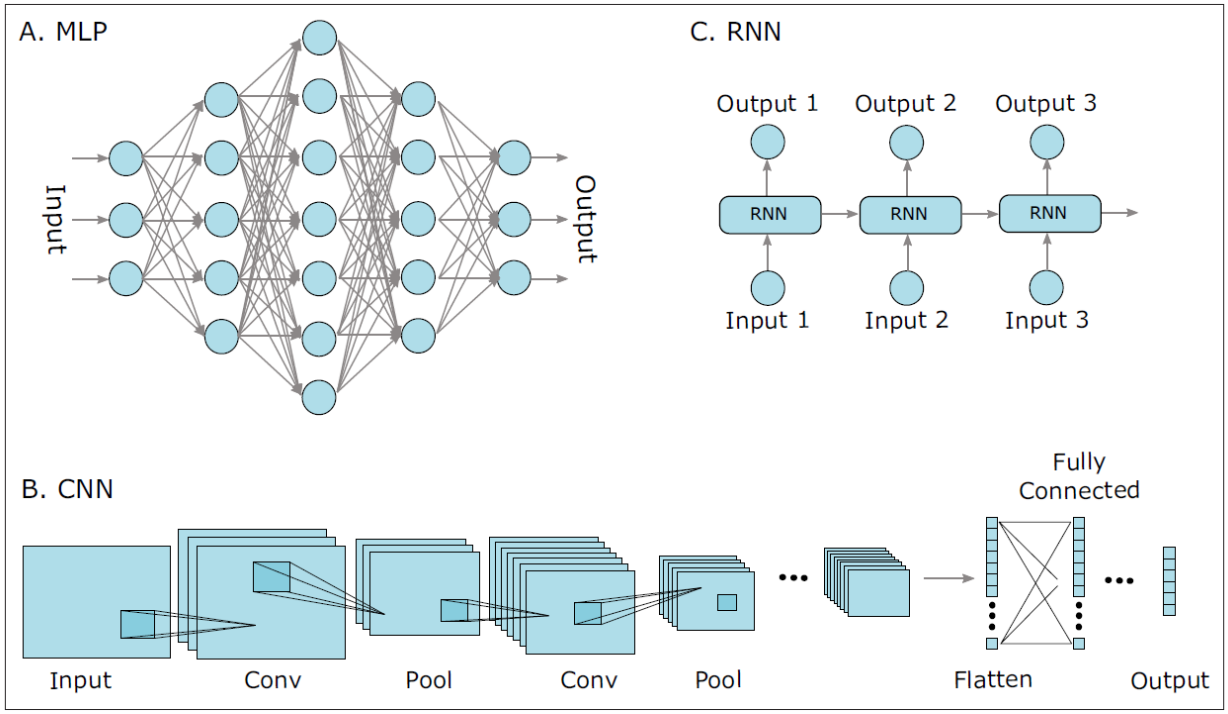


***Figure 3: Three prominent deep learning architectures include**:*

(A) Multilayer Perceptrons (MLPs): These networks consist of an input layer, several hidden layers, and an output layer. (B) Convolutional Neural Networks (CNNs): In CNNs, convolution operations are used within layers to learn filters that automatically extract features from the input data (e.g., images, audio signals, time series, or protein sequences). Eventually, the learned features are flattened into a vector, which is typically fed into one or more fully connected layers at the end. (C) Recurrent Neural Networks (RNNs): RNNs process input sequences one element at a time, handling the sequence step by step.

RNNs introduce sequential processing capabilities by maintaining an internal state that captures information from previous inputs, making them especially suited to time series and sequence data. This architecture is advantageous in handling the sequential nature of protein data and has been effective in combination with CNNs for hybrid models in protein analysis.

MLPs, CNNs, and RNNs, play a pivotal role in protein subcellular localization prediction. Each architecture brings specific strengths to bioimage analysis, with CNNs excelling in capturing spatial patterns within image data, while RNNs handle sequence data effectively (Chandra et al., 2023). Selecting the appropriate architecture is critical in building effective prediction models, as each type of neural network offers unique advantages depending on the nature of the data and the prediction goals (Koumakis, 2020; Liu et al., 2007; Mikolov et al., 2013b; Zeng et al., 2016).

Integrating DL into protein subcellular localization prediction represents a significant step forward in enhancing both the accuracy and efficiency of classification models, paving the way for a more nuanced understanding of protein function and cellular processes (Zeng et al., 2016).

2.5 CNN for Protein Subcellular Localization Prediction

In recent years, CNNs have emerged as the dominant approach for tasks involving image classification, segmentation, and object detection, particularly in data-rich fields like biological image analysis (Lambin et al., 2014; Aerts et al., 2014; Yamashita et al., 2018; Chandra et al., 2023)). Unlike traditional methods such as Loc-CAT, which rely on hand-crafted features, such as texture or intensity measures, followed by machine learning classifiers like random forests or SVMs, which require significant human intervention in feature selection and workflow design (Lambin et al., 2014; Aerts et al., 2014), CNNs process raw images directly, learning hierarchical feature representations end-to-end. This ability enables CNNs to capture cellular localization patterns with remarkable efficiency and scalability (Liimatainen et al., 2021; Ehteshami et al., 2017; Krizhevsky et al., 2012; Ouyang et al., 2019).

CNNs are now foundational to biological image analysis, with applications ranging from multi-label classification of protein localization in yeast to more complex cellular analyses (Moen et al., 2019; Godinez et al., 2017; Hofmarcher et al., 2019). Key advancements in CNN architectures, such as ResNet (He et al., 2016), Inception (Srivastava et al., 2014), and DenseNet (Huang et al., 2017), alongside training techniques like Dropout, Batch Normalization, Focal Loss, Cyclical Learning Rates, and AutoAugment, have substantially improved CNN performance and accessibility (He et al., 2016; Huang et al., 2017; Srivastava et al., 2014; Ioffe & Szegedy, 2015). Libraries like PyTorch (Paszke et al., 2017) and TensorFlow (Abadi et al., 2016) have made CNNs accessible to a broad audience, while AutoML techniques like hyperparameter optimization, meta-learning, and neural architecture search streamline model development and reduce the need for deep technical expertise (Abadi et al., 2016; Lin et al., 2017; Paszke et al., 2017; Smith, 2017; Falkner et al., 2018; Cubuk et al., 2018; Vanschoren, 2018; Hutter et al., 2019).

2.5.1 Building Blocks of a CNN: Convolution and Convolutional Layers

CNNs are designed for grid-structured data, such as images, drawing inspiration from the organization of the animal visual cortex (Fukushima, 1980; Aerts et al., 2014). CNNs automatically learn spatial hierarchies of features, progressing from simple to more complex patterns, which is key to their efficacy in image processing tasks (Yamashita et al., 2018).

The core architecture of a CNN comprises three primary layer types: convolutional layers, pooling layers, and fully connected layers. The convolutional layers perform feature extraction by applying learnable filters (or kernels) across the input data, detecting specific features in the image through a sliding window approach (Nair & Hinton, 2010; Yamashita et al., 2018). Each kernel generates a feature map, or activation map, which highlights particular characteristics in the input, enabling the network to identify a diverse array of patterns in the data (Nair & Hinton, 2010) (Figure).

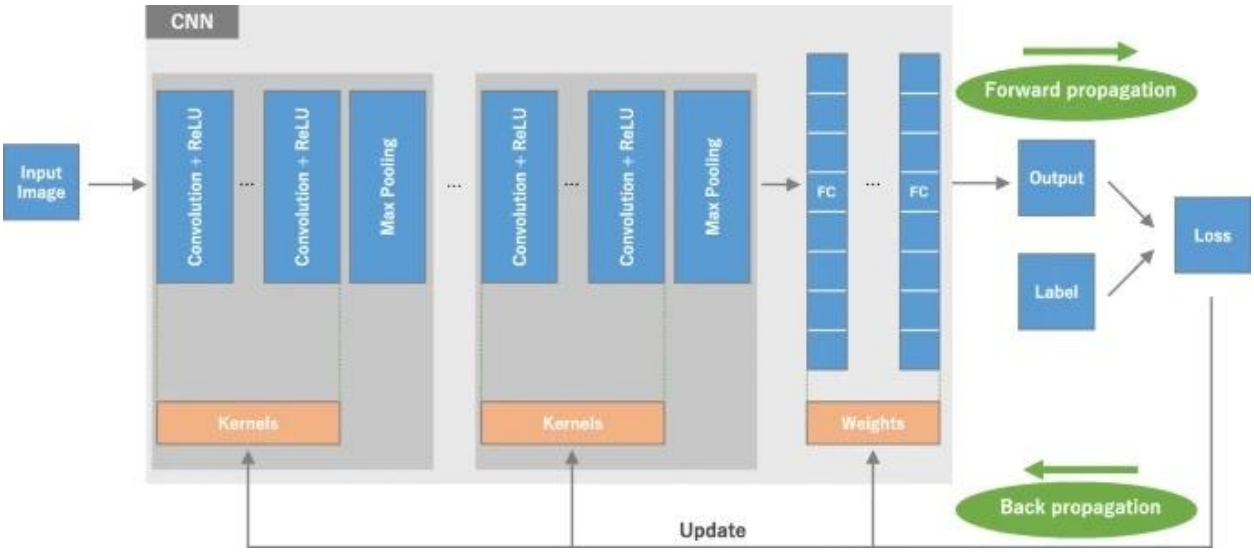


Figure 4: Components of a convolutional neural network (CNN)

A CNN is built by layering several fundamental components, including convolutional layers, pooling layers (such as max pooling), and fully connected (FC) layers. The network's performance, given a specific set of kernels and weights, is assessed by a loss function during forward propagation over a training dataset. Subsequently, these learnable parameters—namely the kernels and weights—are adjusted using backpropagation combined with a gradient descent optimization algorithm. The activation function ReLU (rectified linear unit) is also utilized in this process.

Following the convolutional layers, pooling layers reduce the spatial dimensions of the feature maps, increasing computational efficiency and supporting the network's hierarchical feature

learning process. Pooling preserves the most relevant aspects of each feature map by downsampling, typically using max-pooling, which selects the maximum value within each region (Yamashita et al., 2018). This dimensionality reduction also mitigates overfitting by focusing on the most prominent features and allows the CNN to recognize features invariant to minor transformations.

Finally, fully connected layers aggregate the learned features from convolutional and pooling layers, generating a final prediction, such as classifying an image into predefined categories. This structure enables CNNs to effectively model complex relationships within the data (Glorot et al., 2011; Ramachandran et al., 2017).

2.5.2 Convolution Operations and Network Efficiency

The convolution operation lies at the heart of CNN functionality, involving the application of a kernel across the input tensor, which is often a 2D grid of pixel values in digital images (Ramachandran et al., 2017). The kernel performs element-wise multiplication as it slides over the input, summing the results to produce corresponding values in the output feature map. This process, known as weight sharing, reuses the same kernel across all input regions, ensuring translation invariance—the ability to recognize features regardless of their position in the image. Weight sharing also significantly reduces the model's parameter count, making CNNs more computationally efficient (Ramachandran et al., 2017) (Figure).

The convolution operation described earlier does not permit the center of a kernel to align with the outermost element of the input tensor, resulting in an output feature map that is smaller in height and width than the input tensor. To remedy this, padding—typically zero padding—is applied. This technique involves adding rows and columns of zeros around the input tensor, allowing the kernel's center to cover the outermost elements and maintaining the same spatial dimensions during the convolution process (Yamashita et al., 2018; Koumakis, 2020). Contemporary CNN architectures generally use zero padding to preserve in-plane dimensions, enabling the application of additional layers. (Nair & Hinton, 2010; Glorot et al., 2011; Ramachandran et al., 2017). Additionally, the stride parameter, or step size of the kernel's movement, can be adjusted to control downsampling. While a stride of one is typical, larger strides

can further reduce the feature map dimensions, complementing pooling for efficient feature extraction (Yamashita et al., 2018).

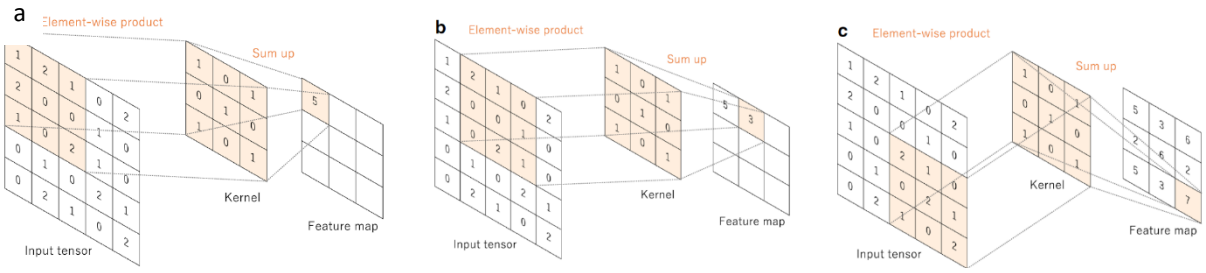


Figure 5: Illustration of convolution operation using a 3×3 kernel

a-c Consider a convolution operation using a 3×3 kernel, no padding, and a stride of 1. In this process, the kernel is systematically moved across the input tensor. At each position, an element-wise multiplication is performed between the kernel and the corresponding section of the input tensor, and the resulting products are summed. This sum forms the output value at the corresponding location in the resulting tensor, known as a feature map.

2.5.3 Optimizing CNNs through Training

Training a CNN involves optimizing the kernel weights to minimize prediction errors relative to the true labels, a process achieved through backpropagation and gradient descent algorithms (Yamashita et al., 2018). In backpropagation, errors are calculated at the output layer and propagated backward, adjusting the weights in each layer to reduce discrepancies in the next iteration. This iterative process allows CNNs to learn highly effective feature extractors specific to the prediction task, which is essential for achieving high performance and accuracy (Yamashita et al., 2018).

CNNs' powerful combination of automated feature extraction, hierarchical learning, and efficiency makes them indispensable in tasks requiring spatial pattern recognition, such as protein subcellular localization prediction. With their ability to learn from large-scale data and adapt to intricate image patterns, CNNs are now pivotal in advancing our understanding of protein localization and function within cellular environments.

2.5.4 Advances in CNN Architectures for Protein Localization Analysis

Modern CNN architecture also extends to three-dimensional data (3D-CNNs), allowing for volumetric processing, which is particularly beneficial in studying spatially complex biological samples (Howard, & Gugger, 2020). For protein localization, 3D-CNNs provide enhanced capability to capture depth-related information, modeling three-dimensional relationships across subcellular structures to yield a more comprehensive view of protein distribution within cells (Yamashita et al., 2018; Koumakis, 2020). 3D-CNNs can deepen our understanding of protein behavior and interactions within the cellular landscape by capturing these spatial hierarchies.

2.6 Transformer Models

The Transformer model, introduced by Vaswani et al. in 2017, initially achieved state-of-the-art results in language translation, notably reducing training times compared to previous models (Vaswani et al., 2017; Chandra et al., 2023). The Transformer is a neural network architecture mainly designed for natural language processing (NLP) tasks, including language translation, text summarization, and question-answering (Pratap, 2023). Its primary breakthrough lies in the self-attention mechanism, which enables the model to assess and prioritize different parts of the input when generating predictions (Pratap, 2023). A Transformer is an encoder-decoder model: the encoder maps input sequences into internal representations, while the decoder interprets these representations to generate output sequences (see Figure 6 A for a general structure). Unlike CNN models, the Transformer uses an architecture component called attention (Table 1), which enables the model to assess interactions between all elements of a sequence and automatically identify relationships crucial for prediction (Cheng et al., 2021; Chandra et al., 2023).

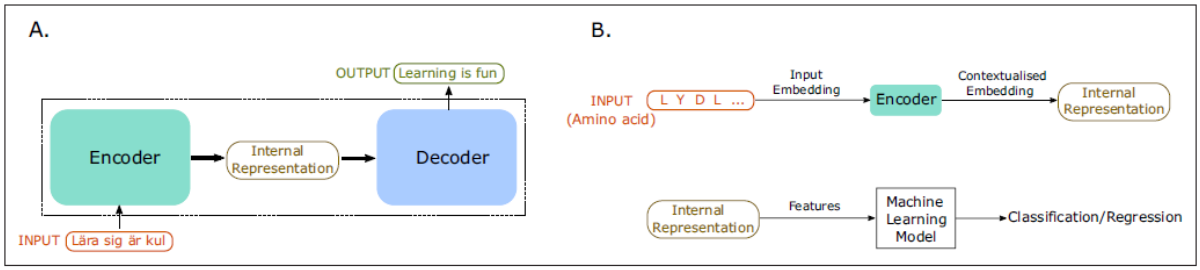


Figure 6: Sequence-to-sequence framework using transformer models

(A) Depicts the conceptual framework of sequence-to-sequence modeling, similar to the Transformer model proposed by Vaswani et al. (2017), which relies on an encoder-decoder architecture to convert an input sequence into an output sequence.(B) Provides

an example of using the Transformer language model to predict protein properties. The encoder block contextualizes the input embedding, creating an internal representation—essentially the model's learned embedding of the input sequence. This representation can be extracted as amino acid features and passed along to a separate machine learning model in a subsequent step. Typically, the decoder block is not employed after training for protein property prediction, since it offers little benefit for that task; however, it remains an essential element in natural language processing (NLP) applications, such as machine translation.

The core innovation in Transformers lies in multi-head self-attention, where multiple attention modules work in parallel to learn various relationships within the input sequence. Each attention head assigns weights to every token relative to others in the sequence, allowing the model to focus on essential parts of the data dynamically (Vaswani et al., 2017). Self-attention mechanisms enable the model to determine the significance of various parts of the input when generating predictions. This process involves calculating a dot product between the query, key, and value vectors for each element in the input sequence. The resulting dot product helps compute a weighted sum of the values, where the weights are assigned based on the similarity between the query and key vectors. This approach allows the model to dynamically focus on relevant parts of the input rather than depending on a predefined set of learned features. This attention mechanism makes Transformers particularly effective in tasks that require understanding both local and global dependencies, such as analyzing spatially complex protein structures (Chandra et al., 2023). Their ability to model complex dependencies between elements in a sequence, such as spatial relationships in protein structures, presents a promising avenue for advancing protein analysis beyond the current capabilities of CNNs.

Table 1: Key Components of CNNs and Transformers

Model Type	Component	Function
CNN	Convolutional Layers	Utilize filters to scan input data and identify distinct features within an image.
	Pooling Layers	Reduce the spatial size of feature maps to enhance model robustness against image translations.

	Fully Connected Layers	Process the extracted features from previous layers to generate predictions.
Transformer	Multi-head Self-Attention Layers	Compute relationships between different parts of the input sequence using query, key, and value vectors, helping the model focus on relevant information.
	Feed-Forward Neural Networks	Further process the self-attention outputs to refine predictions.
	Layer Normalization	Normalizes inputs across layers to improve training stability and efficiency.

The Vision Transformer (ViT) adapts the Transformer model for computer vision tasks by representing images as sequences of patches instead of a pixel grid (Rustamy, 2023; Pratap, 2023). These patches are then processed through multiple layers, similar to the encoder structure in the original Transformer (Pratap, 2023). ViT employs a self-attention mechanism akin to the Transformer but with notable differences. Instead of using convolutional layers like traditional CNNs, ViT applies a linear projection to process image patches, enabling the model to learn a more flexible and expressive feature space (Pratap, 2023). Additionally, ViT incorporates a multi-head self-attention mechanism, utilizing multiple sets of query, key, and value vectors. This allows the model to focus on various aspects of the input simultaneously, enhancing its ability to capture meaningful patterns in visual data (Pratap, 2023).

Transformers are typically trained using self-supervised learning, often through one of two methods: (1) autoregressive language modeling, where the model predicts the next token in a sequence based on previous tokens, or (2) masked language modeling (MLM), where a subset of tokens is masked, and the model learns to predict these tokens using the remaining context (Peters

et al., 2018; Devlin et al., 2018). MLM, in particular, has proven highly successful, as it enables Transformers to consider the entire input sequence, making them adept at capturing long-range dependencies and yielding accurate sequence embeddings (Nambiar et al., 2020; Elnaggar et al., 2020a; Rives et al., 2021; Brandes et al., 2021; Rao et al., 2021; He et al., 2021).

2.6.1 Solving Protein Prediction Tasks Using Transformers

While Transformer models have shown remarkable promise in various protein prediction tasks, their application to protein subcellular localization prediction based solely on image data remains limited (Chandra et al., 2023). This limitation arises from challenges such as the novelty of image-based prediction methods, the complexity of protein localization patterns, and the scarcity of labeled datasets for training large Transformer models (Chandra et al., 2023). However, as Transformers evolve and demonstrate their potential across protein-related tasks, there is growing interest in adapting them for localization prediction within bioimage data.

Transformers have been effectively applied to several protein prediction tasks, including protein structure prediction, residue contact mapping, protein–protein interactions (PPI), drug–target interactions (DTI), post-translational modifications (PTMs), and homology studies. These tasks can be categorized as either local (focusing on specific sites of interest within a sequence) or global (analyzing the entire sequence). For local tasks, a fixed-size Transformer representation can be generated by selecting a fixed window around the sites of interest, whereas for global tasks, a sequence representation is achieved by averaging residue vectors, providing a robust protein sequence vector (Väth et al., 2022).

A key advantage of Transformer models in these tasks is their ability to generate meaningful representations without relying on multiple sequence alignments (MSAs) or structural information. Traditional methods often use profile-to-profile comparisons derived from MSAs, such as PSI-BLAST (Altschul et al., 1997) and HMMER (Finn et al., 2011). However, MSA-based techniques have inherent limitations due to sequence gaps from deletions and insertions (Golubchik et al., 2007) and perform poorly on sequences with few homologs (Phuong et al., 2006). Moreover, although structural features like secondary structures can enhance predictive models, they remain imperfect due to unresolved structural information challenges (Sułkowska et al., 2012; Schmiedel & Lehner, 2019).

Transformers bypass these limitations by creating embeddings directly from sequence data. For instance, recent frameworks such as that by Chowdhury et al. (2022) have used Transformer-based language models to outperform even advanced MSA-based tools like AlphaFold2 (Jumper et al., 2021) on structure prediction tasks for sequences lacking homologs. Some Transformer models also incorporate evolutionary information from MSAs during pre-training, but once pre-trained, they can generate representations for new proteins using only the internal states of the model. This approach not only reduces computational costs but also builds richer and more complete features for proteins with low homology, offering a compelling alternative to traditional MSA tools that require extensive searches across databases like UniProt—a time-intensive process (Hong et al., 2021; Wang et al., 2022).

In protein subcellular localization, ongoing research highlights promising strategies to adapt Transformer-based architectures for image data. Methods such as GraphLoc, which utilizes graph neural networks with image data, demonstrate the feasibility of capturing spatial relationships within protein images, while approaches like DeepLoc apply deep learning techniques directly to image-based protein localization (Chandra et al., 2023). Although these models do not directly incorporate Transformers, they provide foundational steps for exploring Transformer architectures in protein subcellular localization (Chandra et al., 2023). Leveraging the attention mechanisms within Transformers could enable these models to capture complex relationships within bioimages, transcending the limitations of CNNs by considering global contextual dependencies and long-range spatial interactions (Chandra et al., 2023).

Transformers' ability to capture intricate sequence relationships, bypass traditional alignment methods, and generate representations from diverse data types suggests a promising direction for applying them to protein localization tasks (Chandra et al., 2023). As Transformer models continue to be adapted for protein-related analyses, their application in protein subcellular localization may unlock new dimensions of accuracy and efficiency in bioimage analysis (Chandra et al., 2023).

2.7 Transformer models vs CNN models

Unlike conventional DL methods like RNNs, which process sequences step-by-step, Transformers handle sequences holistically. RNNs often suffer from issues such as vanishing and exploding

gradients and are limited by their inability to capture context from both directions within a sequence (Bengio et al., 1994; Pascanu et al., 2013; Hanin, 2018). Furthermore, RNNs cannot efficiently parallelize computations, leading to slower training (Wang et al., 2019). CNNs, while successful for image analysis, rely on hierarchical feature learning through local receptive fields, which can make them less efficient in capturing long-range dependencies within sequential data, as multiple layers are required to extend the receptive field (Raghu et al., 2021). Additionally, CNNs are spatially invariant, meaning they do not leverage positional information within sequences—an aspect that can be critical in protein localization tasks (Albawi et al., 2017).

Transformers address these challenges by utilizing the attention mechanism to consider relationships between all elements in a sequence directly. Each token can influence the weights of all other tokens, allowing the model to attend to distant dependencies without needing extensive layers (Dehghani et al., 2018). This direct interaction between tokens enhances training efficiency and captures nuanced relationships essential for protein analysis, positioning Transformers as a valuable alternative to CNNs.

Moreover, Transformers are highly parallelizable, allowing for efficient computation due to their reliance on attention modules and fully connected layers rather than recurrent or convolutional structures. This makes Transformers not only computationally efficient but also highly scalable for handling complex protein data (Wang et al., 2019; Väth et al., 2022). By integrating both local and global information, Transformers offer an advanced approach to subcellular localization tasks, potentially unlocking more intricate insights into protein function and organization within cells (Wang et al., 2019).

2.7.1 Compare and Contrast CNNs and Transformers

CNNs and Transformer models bring unique strengths to the protein subcellular localization prediction field, each excelling in specific domains. CNNs are particularly effective at extracting spatial features from protein images, with their hierarchical structure allowing them to recognize detailed patterns and spatial hierarchies. This strength makes CNNs well-suited for tasks that rely heavily on localized image features. On the other hand, Transformers are powerful in sequence modeling, capturing intricate dependencies and contextual information across entire sequences, which is invaluable in modeling complex relationships within protein data.

A promising area of innovation lies in hybrid models that combine the spatial feature extraction of CNNs with the sequence modeling capabilities of Transformers. These hybrid models could leverage the complementary strengths of both architectures, enabling them to capture the nuanced interplay between protein image features and sequence information. This synergy holds the potential to significantly improve the accuracy, robustness, and interpretability of subcellular localization predictions.

2.7.2 Handling Spatial Dependencies in Transformers

Transformers rely on self-attention mechanisms to model spatial dependencies in images, which differs fundamentally from the convolution-based operations in CNNs. In a CNN, each convolutional filter focuses on local receptive fields, capturing patterns in a patchwise manner. To learn larger spatial contexts, CNNs stack multiple layers so that the receptive field gradually expands (Rustamy, 2023; Pratap, 2023). In contrast, Vision Transformers (ViTs) split an image into patches, project each patch into a latent embedding, and then apply multi-head self-attention across all patch embeddings simultaneously (Liu, Qian, Xia, & Wang, 2024). This attention-based approach enables ViTs to aggregate information from any part of the image, regardless of the physical distance between patches. As a result, Transformers can capture global dependencies in a single step, whereas CNNs typically require deeper architectures and pooling operations to learn long-range relationships (Liu, Qian, Xia, & Wang, 2024).

Because Transformers treat spatial positions more explicitly through positional embeddings, they can better preserve the absolute or relative locations of image regions (Liu, Qian, Xia, & Wang, 2024). However, this global attention can be computationally more expensive since it considers every patch's relationship with every other patch. Nonetheless, for tasks like protein subcellular localization—where fine-grained details and overarching spatial context can both be pivotal—Transformers may offer an advantage by modeling complex spatial interactions more directly than CNNs (Liu, Qian, Xia, & Wang, 2024) .

2.7.3 Key Advantages and Disadvantages of Transformers Over CNNs in Image Analysis

Transformers offer several notable benefits for image analysis. First, their global contextual awareness allows them to capture relationships between distant regions of an image in a single forward pass, which can be especially useful for biological images where important signals may be scattered across the cell. Next, because images are treated as sequences of patches, Transformers exhibit greater flexibility with varying image sizes, bypassing the rigid kernel structures typically found in CNNs (Liu, Qian, Xia, & Wang, 2024). Furthermore, self-attention mechanisms enable substantial parallelization, as computations between patches remain independent until attention weights are computed. This can accelerate training on large datasets, provided there is sufficient computational power (Liu, Qian, Xia, & Wang, 2024). Finally, Transformers embody a reduced inductive bias, relying purely on the data to learn spatial dependencies rather than assuming local invariance, potentially capturing subtle or complex patterns that might elude convolutional filters (Liu, Qian, Xia, & Wang, 2024).

Despite their strengths, Transformers also have limitations. Their quadratic computational complexity can impose a bottleneck when dealing with high-resolution images or large batch sizes, leading to elevated processing costs (Roell, n.d; Liu, Qian, Xia, & Wang, 2024) . Additionally, Transformers tend to be data-hungry, often requiring extensive labelled datasets or self-supervised pretraining to perform effectively (Roell, n.d; Liu, Qian, Xia, & Wang, 2024). This dependence can result in overfitting for smaller datasets unless substantial regularization or data augmentation is employed. Another challenge is the need for explicit positional embeddings to encode spatial information, while CNNs inherently capture locality through convolutions (Roell, n.d; Liu, Qian, Xia, & Wang, 2024) . Finally, interpreting Transformer-based filters can be more challenging than their CNN counterparts, as attention maps—though powerful—may demand more advanced visualization techniques to yield clear biological insights (Roell, n.d; Rustamy, 2023; Pratap, 2023; Liu, Qian, Xia, & Wang, 2024).

2.8 Future Directions

Looking forward, several key directions stand out for advancing deep learning in protein localization prediction. Developing hybrid CNN-Transformer architectures represents a critical

next step, as these models could bridge the gap between image-based and sequence-based analysis, ultimately creating a more holistic approach to protein localization (Raghu et al., 2022; Pratap, 2023). Additionally, graph neural networks (GNNs) offer a promising pathway for protein prediction tasks by capturing relationships within protein structures and across cellular environments, further enhancing predictive capabilities (Fout et al., 2017; Gainza et al., 2020; Zhang et al., 2023).

Moreover, as DL models evolve, there is a growing need for interpretable models to elucidate the biological mechanisms driving protein localization patterns (Lomenick et al., 2011; Wang & Wei, 2022). Interpretability will be essential for advancing scientific understanding and fostering trust in model predictions, particularly in applications related to health and disease,

Multi-modal data integration and the creation of large-scale, high-quality annotated datasets will also be pivotal in increasing the generalizability of prediction models. Integrating diverse data sources—such as genomic, proteomic, and imaging data—can provide a more comprehensive view of protein behavior, thus enhancing model performance and applicability across varied biological contexts.

Ultimately, these future endeavors hold the potential to revolutionize biological research and open new doors in drug discovery and personalized medicine, offering precise insights into protein function, localization, and their implications for human health.

2.9 Rationale of the Study

Although CNN-based approaches have demonstrated high accuracy in predicting protein subcellular localization from bioimages, they often lack interpretability regarding which features drive classification decisions (Jiang et al., 2021). Meanwhile, Transformer-based architectures, which excel in modelling long-range dependencies, remain relatively underexplored for protein localization tasks (Dosovitskiy et al., 2020). Addressing this gap is essential for determining whether these novel attention-driven methods can outperform or complement CNNs in accuracy, scalability, and interpretability (Dosovitskiy et al., 2020). Furthermore, as high-throughput imaging and large-scale proteomics datasets continue to grow, a systematic comparative analysis will help clarify which modelling techniques are most suitable for real-world applications (Wang & Wei, 2022; Xiao et al., 2024). By investigating both CNN- and Transformer-based pipelines

under a unified experimental framework, this study aims to provide new insights into the design of robust, explainable models. This work also improves upon previous studies by integrating carefully curated immunofluorescence microscopy data, exploring both local feature extraction (CNNs) and global contextual modeling (Transformers), and thus laying the groundwork for more precise protein localization predictions that could accelerate discoveries in cell biology and drug development (Ouyang et al., 2019; Jiang et al., 2021; Liimatainen et al., 2021). The accurate subcellular localization of proteins is fundamental in understanding cellular function, disease mechanisms, and drug discovery (Lomenick et al., 2011; Wang & Wei, 2022). Mislocalized proteins have been implicated in a wide range of diseases, including neurodegenerative disorders (e.g., Alzheimer's, Parkinson's), metabolic syndromes, and various cancers, where their aberrant distribution can disrupt normal cellular processes (Lomenick et al., 2011; Wang & Wei, 2022). In biomedical and systems biology research, subcellular localization is critical for identifying therapeutic targets, understanding protein-protein interactions, and elucidating pathways involved in disease pathogenesis (Garapati et al., 2020; Wang & Wei, 2022). Pharmacologically, predicting protein localization can enhance drug target identification, optimize therapeutic interventions, and improve precision medicine approaches by revealing protein behavior within different cellular compartments (Garapati et al., 2020; Wang & Wei, 2022).

Traditionally, protein localization has been studied using experimental techniques such as immunofluorescence microscopy, mass spectrometry-based proteomics, and biochemical fractionation (Jakobsen et al., 2011; Christoforou et al., 2016; Itzhak et al., 2016; Orre et al., 2019). While these methods provide high-resolution insights into protein distributions, they are labour-intensive, costly, and often unsuitable for large-scale applications (Xiao et al., 2024). Consequently, computational approaches have emerged as efficient alternatives, leveraging bioimage data and predictive modelling to infer subcellular localization from large-scale datasets (Chebira et al., 2007, Hamilton et al., 2007, Khan et al., 2008, Khan et al., 2011, Lin et al., 2007, Murphy et al., 2003, Nanni and Lumini, 2008, Nanni et al., 2010a, Zhang et al., 2009).

With the rise of AI and deep learning, significant advancements have been made in automating and accelerating protein localization prediction (Xiao et al., 2024; Coursera Staff, 2024). Deep learning models, particularly CNNs, have demonstrated remarkable accuracy in analysing bioimages by capturing intricate spatial patterns. CNN-based approaches have successfully

identified protein locations using immunofluorescence images, enabling large-scale analyses that would be impractical using traditional methods (Liimatainen et al., 2021; Ehteshami et al., 2017; Krizhevsky et al., 2012; Ouyang et al., 2019). However, despite their high accuracy, CNNs have notable limitations, particularly in learning long-range dependencies and providing explainable predictions (Gai et al., 2024). CNNs primarily focus on local spatial relationships, making them less effective at capturing complex global dependencies within cellular structures (Gai et al., 2024).

Recent developments in AI have introduced Transformer-based architectures, which have revolutionized sequence-based tasks, notably in natural language processing (e.g., BERT, GPT) and protein structure prediction (e.g., AlphaFold) (Jumper et al., 2021). AlphaFold, developed by DeepMind, has significantly accelerated biological discovery by accurately predicting protein 3D structures, surpassing previous computational models and experimental techniques in speed and precision (Roell, n.d; Rustamy, 2023; Pratap, 2023; Liu, Qian, Xia, & Wang, 2024). Given the success of Transformers in modelling complex relationships, their potential in subcellular protein localization remains underexplored (Barmada et al., 2010; Lomenick et al., 2011; Wang & Wei, 2022; Ziff et al., 2023). Unlike CNNs, which rely on local feature extraction, Transformers utilize self-attention mechanisms to capture both local and global dependencies, making them particularly suited for analysing high-dimensional, structured bioimage data (Casola, Lauriola & Lavelli, 2022; Chen, 2022; Pratap, 2023). However, Transformers have limitations as well (Casola, Lauriola & Lavelli, 2022). Their computational complexity makes them expensive to train, and they often require very large datasets to generalize effectively (Casola, Lauriola & Lavelli, 2022; Chen, 2022; Pratap, 2023). These constraints raise an important question: **Can Transformers outperform or complement CNNs for protein localization, particularly in real-world bioimage datasets that may not always be large or well-annotated (Gai et al., 2024)?**

This study aims to bridge this gap by evaluating both CNN- and Transformer-based models for protein subcellular localization prediction. By leveraging high-throughput immunofluorescence microscopy datasets, we systematically compare these two deep learning paradigms within a unified experimental framework. Specifically, CNNs excel at capturing fine-grained spatial patterns, while Transformers offer enhanced scalability and global context modelling (Liimatainen et al., 2021; Ehteshami et al., 2017; Krizhevsky et al., 2012; Ouyang et al., 2019). By implementing

and optimizing both approaches, we aim to determine their relative strengths, trade-offs, and potential synergies in protein localization tasks.

Furthermore, this study seeks to enhance model interpretability by employing visualization techniques such as Grad-CAM for CNNs and attention rollouts for Transformers (Selvaraju et al., 2017b; Moujahid et al., 2021). This will provide insights into how these models identify subcellular structures and contribute to the development of explainable AI in bioinformatics(Selvaraju et al., 2017b; Moujahid et al., 2021) .

Beyond its computational contributions, this research carries significant implications for experimentalists working in molecular and cell biology. High-throughput protein localization analysis is a crucial yet time-consuming task, and deep learning-based tools can dramatically reduce the reliance on labour-intensive experimental methods. By identifying the most suitable deep learning framework for subcellular protein localization, this study aims to provide computational biologists and experimental researchers with more efficient, scalable, and interpretable tools for large-scale protein analysis, ultimately reducing the time and cost of experimental validation. These insights can also facilitate large-scale screening of novel protein functions and accelerate biomedical discoveries by integrating AI-driven predictions with traditional experimental workflows.

Ultimately, our research contributes to the growing intersection of AI and cell biology by advancing computational techniques for protein localization (Gai et al., 2024). The findings have the potential to accelerate discoveries in molecular biology, facilitate high-throughput screening of novel drug targets, and improve our understanding of disease mechanisms through AI-driven localization predictions. By systematically evaluating CNNs and Transformers, this work lays the foundation for more precise, scalable, and interpretable models in bioimage-based protein localization.

2.10 Objectives

2.10.1 General Aim

To train convolutional neural network (CNN) architectures and Transformer-based models for the multi-label classification of protein subcellular localization in eukaryotic cells using large-scale

immunofluorescence image datasets and subsequently compare their performance to determine their relative strengths and limitations.

2.10.2 Specific objectives

1. To develop and train CNN-based models using immunofluorescence images from the Human Protein Atlas to establish baseline accuracy and F1-scores.
2. To train and optimize Transformer-based models for subcellular localization, assessing their performance under similar training conditions.
3. To assess the explainability of trained models using Grad-CAM for CNNs and attention maps for Transformers, evaluating how each architecture identifies biologically relevant regions in cellular images.
4. To assess model scalability and generalization by examining performance across abundant and underrepresented classes, highlighting strengths and limitations for practical applications in proteomics.

2.10.3 Ethics Statement

This study was approved by the **Faculty of Health Sciences Human Research Ethics Committee (HREC) at the University of Cape Town (UCT)** under ethics clearance number **HREC REF: 557/2023**. This study utilizes publicly available data from established repositories, specifically the Human Protein Atlas (HPA). All images and metadata from HPA are collected under protocols approved by relevant institutional review boards, and the data are shared with the research community for non-commercial purposes. No new patient samples or personal identifying information were collected for this study. Hence, additional ethical approval was not required. All analyses and results comply with the terms of use stipulated by the data providers.

Chapter Three- Methods

3 Data Collection and Dataset Preparation

All experiments were conducted using a high-performance computing environment equipped with a Quadro RTX 8000 GPU (16 GB VRAM) and 64 GB of system RAM. We sourced our dataset from the Human Protein Atlas (HPA) (The Human Protein Atlas, n.d.), focusing on subcellular localization annotations of various proteins (Thul et al., 2017). The data was provided in a tab-separated values (TSV) file containing fields such as 'Gene', 'Gene name', subcellular location reliability categories ('Enhanced', 'Supported', 'Approved', 'Uncertain'), and specific subcellular localization terms. We imported this dataset into a Pandas DataFrame to facilitate preprocessing and exploratory data analysis. All data preprocessing and analysis were performed using Python 3.9 (Van Rossum & Drake, 2009) in a Jupyter Notebook environment (version 7.0.8).

To provide a clear overview of the full workflow, **Figure 7** summarizes the major steps undertaken in this study — from data collection and preprocessing to model training, evaluation, and prediction. This diagram is intended to guide the reader through the methodological framework described in detail in the subsequent sections.

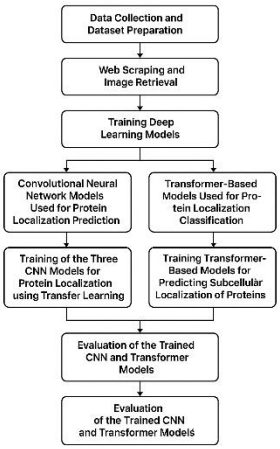


Figure 7: Protein Localization Prediction Workflow.

Overview of the methodological pipeline used for protein subcellular localization prediction. The workflow begins with data collection and preparation, followed by image retrieval, preprocessing, and training of both CNN and Transformer-based models. After model evaluation, final predictions on protein localization are generated

3.1 Web Scraping and Image Retrieval

To obtain the necessary protein images, we programmatically retrieved them from the Human Protein Atlas (HPA) using a Python-based web scraping pipeline. For each protein entry in the dataset, we constructed URLs by combining the base HPA URL (https://www.proteinatlas.org/) with the corresponding Gene and Gene name fields, followed by the /antibody#ICC suffix to access the immunocytochemistry section.

We used the requests library to fetch HTML content from each constructed URL, and parsed the content using BeautifulSoup v4.12.2 (Richardson, 2007). Within each page, we identified <a> tags containing the class color box, and extracted href attributes containing the _selected substring, which indicated high-resolution image links. We modified these links by removing _medium from the path and prepended the base URL to construct the full image download links.

Downloaded images were subjected to a quality control step using a brightness threshold. Images were opened using the Pillow v9. 5.0 library and converted into NumPy v1.23.5 arrays to assess the mean pixel intensity of the green channel, which encodes the localization pattern of the protein of interest. Images with a mean intensity below 70 were excluded to eliminate low-quality or underexposed samples (Rahadianti et al., 2021; Cham, 2023; Rasheed et al., 2023).

Each image was composed of four fluorescent channels: green (protein of interest), blue (nucleus counterstained with DAPI), red (microtubules labeled with anti-tubulin), and yellow (endoplasmic reticulum).

After image retrieval and filtering, we standardized the subcellular localization annotations. For each protein, we merged all non-null annotations across reliability categories (Enhanced, Supported, Approved, and Uncertain) into a unified list. We normalized the names to ensure consistency with our label dictionaries (e.g., mapping 'Rods & Rings' to 'Rods & rings') and converted them into numerical format using predefined mappings.

To prepare the dataset for model training, we implemented a multi-label stratified splitting approach using the MultilabelStratifiedKFold method from the iterstrat v0. 1.7 library (Sechidis et

al., 2011; Trent, 2018). This ensured that class distributions were proportionally maintained across the training, validation, and test sets.

3.2 Training Deep Learning Models

For the classification task, we conducted a comparative study involving both CNNs and transformer-based models. Specifically, we trained three CNN architectures and two transformer-based models to evaluate and compare their performance characteristics. CNNs were implemented using TensorFlow v2.11.0 (Abadi et al., 2016) and Keras v2.11.0 (Chollet, 2015), which provided a high-level, flexible framework for building and fine-tuning convolutional layers. In parallel, we leveraged the Hugging Face Transformers v4.31.0 library (Wolf et al., 2020) to implement and fine-tune Vision Transformer and Swin Transformer architectures. This approach provided a comprehensive assessment of the strengths and limitations of each architectural family in the context of subcellular localization classification, enabling a clearer understanding of how different deep learning frameworks and model designs can capture the spatial and morphological complexities of protein localization.

3.3 Convolutional Neural Network Models Used for Protein Localization Prediction

We employed three convolutional neural network (CNN) architectures—DenseNet121 (Huang et al., 2017), InceptionV3 (Szegedy et al., 2016) , and Xception (Chollet, 2017) —to perform multi-label classification of protein subcellular localization. These three models were originally developed by external research teams and have already been pretrained on the ImageNet dataset, which contains over 14 million images and 1,000 classes (Russakovsky et al., 2015). We leveraged this pretrained foundation through transfer learning by replacing each model's final classification layer with a new fully connected layer consisting of 15 nodes—corresponding to our 15 target classes—and appending a sigmoid activation function. This modification ensures that each class can be predicted independently, thus accommodating the multi-label nature of our task. The revised models were then fine-tuned on our dataset, significantly reducing the amount of data

and computational resources required while retaining critical feature representations from ImageNet(**Error! Reference source not found.**).
We choose these three pretrained CNNs for the following reasons: First, DenseNet121 has demonstrated outstanding performance while requiring less memory and computational power compared to other leading-edge methods and consists of 121 layers organized into densely connected blocks, where each layer is directly connected to all subsequent layers (Huang et al., 2017; Albelwi, 2022). This design enhances gradient flow, mitigates the vanishing gradient problem, and encourages feature reuse, allowing the network to achieve high accuracy with fewer parameters and reduced computational overhead. Its efficiency is advantageous for complex classification tasks that require discerning subtle structural and textural cues (Arulananth et al., 2024). By reusing features from earlier layers, DenseNet121 can effectively learn fine-grained characteristics relevant to subcellular localization (Albelwi, 2022).
Secondly, InceptionV3 comprises 48 layers and incorporates employs Inception modules and factorized convolutions to capture multi-scale features efficiently (Szegdy et al., 2016; Ahmed et al., 2023). The extensive use of batch normalization improves training stability and reduces the overall parameter count while maintaining high accuracy (Dong, Zhao, & Chang, 2020). Pretrained on ImageNet, InceptionV3 has demonstrated its versatility and effectiveness across various domains, including flowers (Maji et al., 2013), apparel (Hsiao et al., 2017), fast foods (Martinel et al., 2016), and traffic signs (Yuan et al., 2018). Its ability to adapt to a wide range of visual tasks makes it well-suited for delineating the intricate patterns of subcellular organelles.

Lastly, Xception (Chollet, 2017; Lo, Yang, & Wang, 2019,) builds upon the Inception architecture by employing depthwise separable convolutions and residual connections (He et al., 2016; Viso 2024) to enhance feature extraction and mitigate representational bottlenecks (Sharma, & Kumar, 2022). It accepts 299×299 RGB inputs and comprises a total depth of 126 layers, with 36 convolutional layers strategically structured into entry, middle, and exit flows (Chen, Yang and Zhang, 2020). These design choices reduce computational complexity while preserving accuracy. By separating channel-wise and spatial operations, Xception reduces computational complexity without sacrificing performance. Its substantial accuracy on ImageNet (Russakovsky et al., 2015) highlights its capability to handle intricate image classification tasks. When fine-tuned for protein subcellular localization, Xception's design supports the nuanced capture of subtle visual distinctions necessary for distinguishing between closely related subcellular compartments.

Model: "functional_4"

Layer (type)	Output Shape	Param #
input_layer_8 (InputLayer)	(None, 299, 299, 3)	0
xception (Functional)	(None, 2048)	20,861,4…
dense_12 (Dense)	(None, 1024)	2,098,176
batch_normalization_12 (BatchNormalization)	(None, 1024)	4,096
dense_13 (Dense)	(None, 1024)	1,049,600
batch_normalization_13 (BatchNormalization)	(None, 1024)	4,096
dense_14 (Dense)	(None, 15)	15,375

*Figure 8**: Modified Xception model summary for multi-label classification***

Shows the model summary of the Xception architecture after replacing the final classification layer with a fully connected layer of 15 nodes and a sigmoid activation function. This modification accommodates multi-label classification of the 15-target protein subcellular localization classes

3.4 Training of the Three CNN models for Protein Localization using Transfer Learning

We resized the Images to meet the input requirements of each model. For DenseNet121, images were resized to 224 × 224 pixels (Albelwi, 2022). For Xception and InceptionV3, images were resized to 299 × 299 pixels (He et al., 2016; Viso 2024). Each original image is subjected to random augmentations every time it is passed through the training pipeline (Hasan et al., 2021). Over 10 epochs of training, each image were effectively augmented 10 times, resulting in 10 distinct variations per original image data augmentation techniques, including random rotations, flips, and zooms, were applied to enhance model robustness and mitigate overfitting (**Error! Reference source not found.**).

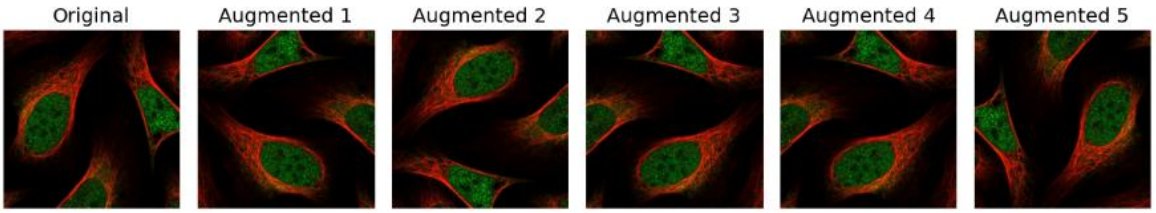


*Figure 9**: Example image augmentations during training***

The first panel shows the original input image, and subsequent panels show variations produced by random rotations, flips, and other transformations applied during training. This illustrates how the model "sees" slightly different versions of the same image each epoch to improve robustness and reduce overfitting

We trained The CNN models were using a batch size of 16 for up to 60 epochs. We then employed the Adam optimizer with an initial learning rate of 1×10^{-4} and then used binary cross-entropy as the loss function due to the multi-label classification nature of this task. Early stopping was implemented to prevent overfitting, monitoring the validation loss with a patience of 10 epochs.

3.5 Transformer-Based Models Used for Protein Localization Classification

In addition to the CNN architectures described above, we explored transformer-based models for the classification of protein subcellular localization. Transformers, developed initially for natural language processing tasks (Vaswani et al., 2017), have increasingly gained prominence in computer vision due to their capacity for flexible, global context modeling without the inherent locality constraints of convolutional filters. Importantly, transformer-based models operate on sequences of image patches rather than entire images, enabling them to capture long-range dependencies and integrate feature representations across spatial locations (Dosovitskiy et al., 2020).

For this study, we implemented and fine-tuned two prominent transformer-based architectures on our dataset: the Vision Transformer (ViT) (Dosovitskiy et al., 2020) and the Swin Transformer (Liu et al., 2021). Similar to the CNN models, both transformers were pretrained on ImageNet (Russakovsky et al., 2015), which facilitated transfer learning. We adapted the final classification layers of these models by introducing a fully connected layer with 15 output nodes—corresponding to our 15 classes—followed by a sigmoid activation function, to handle the multi-label nature of the task.

Briefly, the ViT architecture applies the standard transformer design, initially conceived for sequential data (Vaswani et al., 2017), directly to image patches (Dosovitskiy et al., 2020). Input images are divided into fixed-size patches, which are then linearly embedded and combined with

positional embeddings to retain spatial information. Multi-head self-attention layers allow ViT to attend to different parts of an image simultaneously from multiple representational subspaces, effectively integrating global contextual information (Dosovitskiy et al., 2020; Vaswani et al., 2017). Residual connections (He et al., 2016) and layer normalization (Ba et al., 2016) help stabilize training and mitigate vanishing gradients. Notably, ViTs have demonstrated competitive performance with CNNs in various image classification tasks (Dosovitskiy et al., 2020), and their integrated saliency maps offer interpretability advantages by highlighting regions of interest in the image (Chefer et al., 2021). This global receptive field and robust feature integration can be advantageous for protein subcellular localization, particularly when distinguishing classes with subtle morphological differences.

The Swin Transformer refines the transformer paradigm for computer vision tasks through a hierarchical structure and shifted windowing strategy (Liu et al., 2021). Unlike ViT, which treats images as a flat sequence of patches, the Swin Transformer organizes patches into non-overlapping windows and applies local self-attention within these windows. Periodically shifting the window partitioning achieves greater receptive field coverage while maintaining computational efficiency. This hierarchical approach mirrors the multi-scale feature representation in CNNs, enabling Swin Transformers to capture local and global patterns. The architectural innovations in Swin improve latency and computational efficiency and enhance performance on a range of vision tasks, from image classification to object detection and segmentation (Liu et al., 2021). When applied to our protein subcellular localization problem, the Swin Transformer's hierarchical, windowed attention mechanism can potentially focus on relevant structural details of subcellular compartments while maintaining a global context.

3.6 Training Transformer-Based Models for Predicting Subcellular Localization of Proteins

We employed a slightly different training strategy for the transformer-based models to account for their unique input representations and architectural requirements. While both the Vision Transformer (ViT) (Dosovitskiy et al., 2020) and the Swin Transformer (Liu et al., 2021) ultimately require inputs of images of size 224 × 224 pixels—similar to the CNNs—the data

preprocessing and model compilation procedures were adapted to better align with the transformer frameworks.

First, we integrated the Hugging Face Transformers (Wolf et al., 2020) library, which provided specialized feature extractors for each transformer model architecture. Rather than relying solely on standard TensorFlow/Keras (Chollet, 2015; Abadi et al., 2016) image preprocessing pipelines, images were passed through a feature extraction process that converted them into pixel values tailored for the transformer input format. This step ensured that the multi-head self-attention mechanisms within the models could effectively capture spatial relationships between image patches.

The training regime for the transformer-based models was designed to maintain consistency with the CNN approach while accommodating the transformers' different computational patterns. As with the CNNs, we trained the ViT and Swin Transformer using a batch size of 16 and up to 60 epochs. The Adam optimizer (Kingma and Ba, 2015), initialized with a learning rate of 1×10^{-4}, and binary cross-entropy loss were retained for direct comparability. We also applied similar data augmentation techniques to ensure a fair evaluation, though the feature extraction step was unique to the transformer pipelines (Hasan et al., 2021).

One key difference in the Swin Transformer training was including a hierarchical window-based attention mechanism, which required a custom training argument setup (Liu et al., 2021). Unlike the standard CNN training loop, where the optimizer and learning rate scheduling were managed directly through Keras, the Swin Transformer was compiled using Transformers create_optimizer, allowing for fine-grained control over learning rate schedules and warm-up steps (Kingma and Ba, 2015). This approach aimed to ensure efficient convergence given the increased complexity of transformer-based architectures.

3.7 Evaluation of Model Performance the Trained CNN and Transformer Models

Model performance for both CNN and transformer models was evaluated on the test set comprising 2,473 images. We computed metrics including the accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC) to assess classification

performance. Confusion matrices and ROC curves were generated to visualize the models' performance across different classes.

To ensure a fair comparison, all models were trained and evaluated under the same conditions, using identical training and validation splits and consistent data preprocessing and augmentation techniques.

3.8 Class Activation Mapping and Attention Visualization

To interpret the spatial saliency and attention within our deep learning models, we employed a combination of Grad-CAM–based visualization for convolutional networks and attention-based methods for Transformer architectures.

We utilized Grad-CAM v0.8.1 (Selvaraju et al., 2017) to generate class activation maps (CAMs) for CNNs. We defined a Grad-CAM function that constructs a modified model with an output at the final convolutional layer (i.e., “conv5_block16_concat” for DenseNet121, “block14_sepconv2_act” for Xception, and “mixed10” for InceptionV3). A gradient tape then records the activations and gradients, enabling class-specific localization maps to be computed (Selvaraju et al., 2017). The resulting heatmaps were superimposed on the original images using OpenCV’s color-mapping capabilities (Bradski, 2000). This overlay offers a visual representation of the region's most responsible for a given prediction. For display, we used Matplotlib (Hunter, 2007) to arrange both raw images and heatmaps into a grid layout.

For hierarchical Transformer architectures such as Swin Transformer (Liu et al., 2021), we retrieved raw local window attentions by enabling output_attentions=True in the Hugging Face Transformers framework (Wolf et al., 2020). We employed a TensorFlow-converted Swin-Tiny model, pretrained on ImageNet. Each input image was resized to 224×224 pixels (matching the Swin configuration) and normalized using the AutoImageProcessor class (Cheng et al., 2021). We then selected specific layers and heads (e.g., final stage, head 0) to visualize the attention matrices in local windows. By upscaling and color-mapping these attention values via OpenCV, we obtained heatmaps that highlight the spatial regions of focus for the Transformer’s final attention stage (Bradski, 2000; Cheng et al., 2021).

For Vision Transformer models (Dosovitskiy et al., 2021), we adapted an attention rollout procedure (Abnar & Zuidema, 2020) to aggregate attention across multiple layers. We employed a pretrained “ViT-Base” model from Hugging Face Transformers with output_attentions=True (Wolf et al., 2020). Each image was again resized to 224×224 and processed via a ViT-specific feature extractor. To visualize global attention (including the class token), we computed a cumulative propagation of attention across layers. The final aggregated matrix was then reshaped, normalized, and mapped back onto the original image to highlight the most strongly attended patches (Chefer, Gur, and Wolf, 2020).

By combining Grad-CAM, local window attention maps, and attention-rollout techniques, we obtained a comprehensive view of feature importance for both convolutional and Transformer-based architectures.

Chapter Four – Results

4 Results

4.1 Data Collection and Curation

Following the retrieval and preprocessing of images from the Human Protein Atlas, we evaluated the composition of the dataset to assess its suitability for model training. Initially, the dataset contained 28 unique subcellular localization classes. However, this distribution was highly imbalanced, with certain classes such as "Rods & Rings" and "Cytoplasmic Bodies" comprising fewer than 100 samples, while others like "Nucleoplasm" and "Cytosol" were significantly overrepresented (Figure 10).

To ensure sufficient representation for each class during model training, we applied a filtering step that removed all classes with fewer than 200 instances. This reduced the dataset to a total of 12,565

images across 15 unique subcellular localization classes. The filtered class distribution is illustrated in Figure 9.

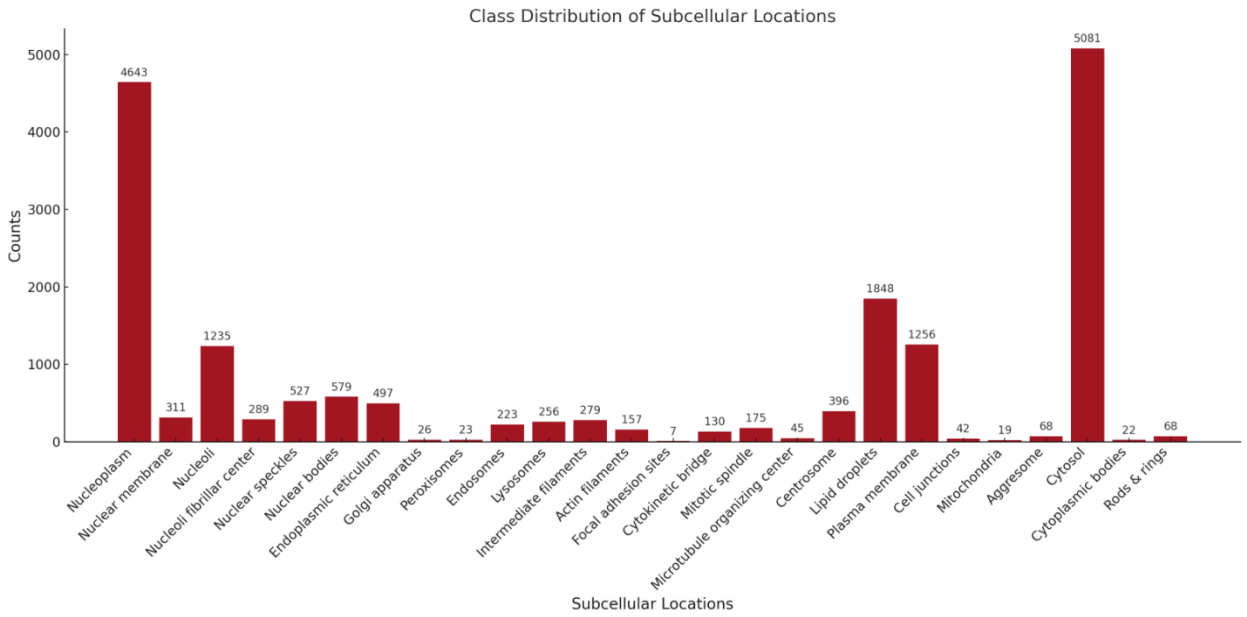


*Figure 10**: Class distribution of subcellular locations in the dataset***

The bar chart illustrates the distribution of instances across various subcellular locations in the dataset. Each bar represents the count of instances for a specific class, labeled on the x-axis. The y-axis shows the number of samples per class. Classes such as "Nucleoplasm" and "Cytosol" have the highest representation, with counts of 4643 and 5081, respectively. In contrast, classes like "Rods & Rings" and "Cytoplasmic Bodies" are underrepresented, with counts below 100

We then prepared the data for model development using a multi-label stratified splitting strategy, appropriate for the nature of subcellular localization annotations which often include multiple labels per image. We employed the MultilabelStratifiedKFold function from the iterstrat library to ensure proportional representation of each class across the training, validation, and testing datasets. Of the 12,565 images, 80% (10,092 images) were allocated to the training set, and 20% (2,473 images) to the test set. Within the training set, 10% (1,009 images) were further set aside for validation. Figure 11 shows the relative percentage distribution of each class label across the entire dataset, training set, and validation set. Notably, the consistency of class proportions across splits indicates a well-balanced stratification.

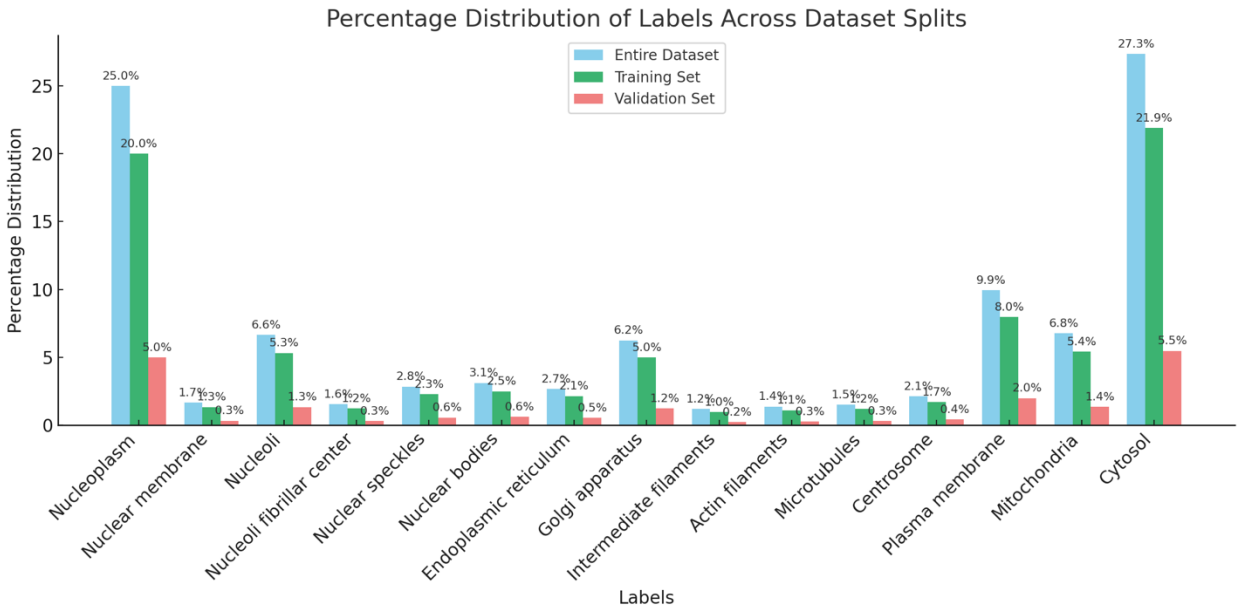


Figure 11: ***Percentage label distribution across data splits***

This bar plot shows the percentage distribution of class labels in the entire dataset (sky blue), training set (medium sea green), and validation set (light coral). Each bar represents the proportion of a specific class, listed along the x- axis, across the respective dataset splits. The y-axis indicates the percentage distribution. The plot highlights the over-representation of certain major classes (e.g., Cytosol, Nucleoplasm) and the under-representation of rare classes (e.g., Nuclear membrane, Intermediate filaments). The consistency between training and validation sets reflects a balanced data split for model training and evaluation. Numerical percentages are annotated above each bar for clarity

To provide a qualitative overview of the dataset, we selected representative microscopy images for each of the 15 valid subcellular localization classes (Figure 12). These examples highlight the distinct morphological features associated with each compartment. Each image includes four fluorescent channels: the green channel marks the protein of interest, while the blue, red, and yellow channels serve as references for the nucleus (DAPI), microtubules (tubulin), and endoplasmic reticulum, respectively. These visual examples support the interpretation of downstream classification results and reinforce the biological relevance of the class labels used.

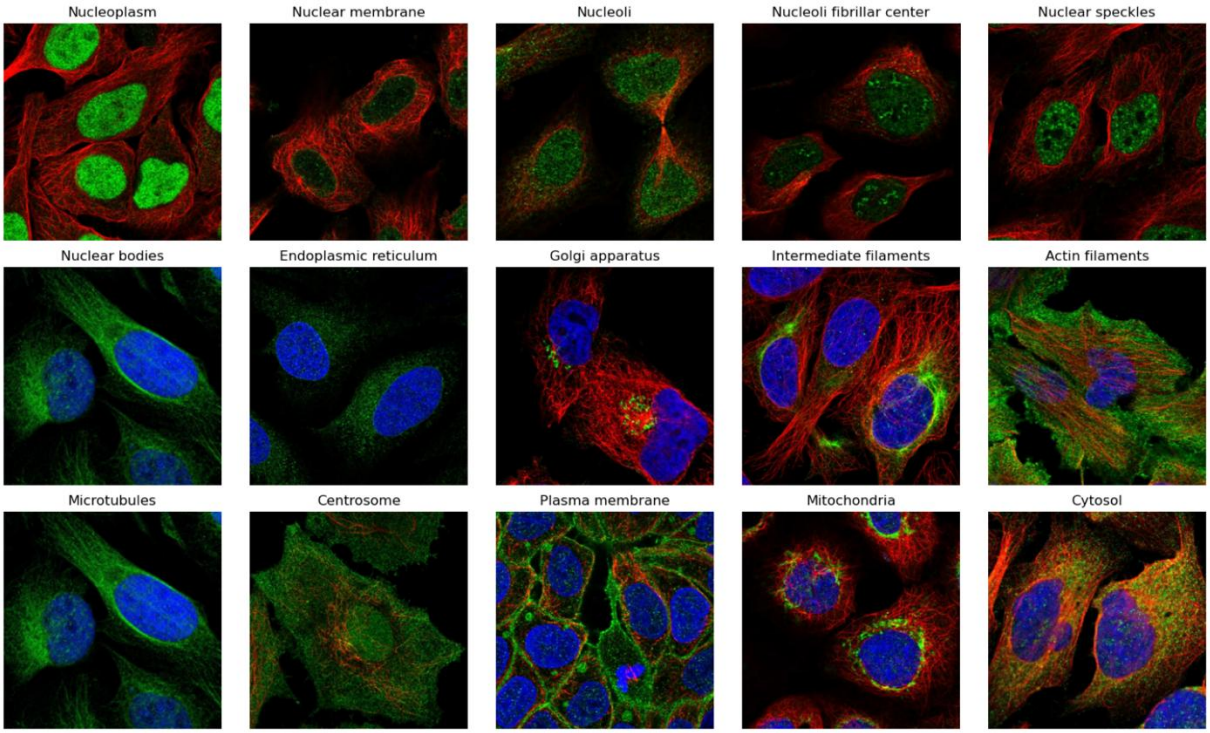


*Figure 12**: Representative images of 15 valid subcellular localization classes***

Representative images illustrating the distinct subcellular localizations of protein we sought to predict in this study. Each panel displays a single cell image chosen from the dataset, exemplifying one of the valid classes included in the analysis (e. g., Nucleoplasm, Nuclear membrane, Golgi apparatus, etc.). The labeling and color channels highlight specific proteins and cellular structures, allowing clear visualization of morphological differences and the unique distribution of markers within each compartment. These representative examples provide a qualitative overview of the cellular features that define each subcellular category and support the quantitative classification results presented

4.2 Overall Models' Performance for Classification of Protein Localization

Table 2. Validation loss and accuracy for each model.

Model	Validation Loss	Validation Accuracy (%)	Test Accuracy (%)
DenseNet121	0.1452	68.62	67.32
Xception	0.1406	66.28	66.14
InceptionV3	0.3294	56.58	58.1
Swin Transformer	0.2196	60.51	61.58
Vision Transformer (ViT)	0.2746	38.0	37.08

We evaluated the performance of three convolutional neural network (CNN) models—DenseNet121, Xception, and InceptionV3—alongside two transformer-based models—the Swin Transformer and the Vision Transformer—on the multi-label classification task of subcellular protein localization using confocal fluorescence microscopy images (Table 2).

Among the CNN models, DenseNet121 demonstrated the best overall performance with a validation accuracy of 68.62%. Although Xception achieved a slightly lower validation loss (0.1406) than DenseNet121 (0.1452), its validation accuracy was marginally lower at 66.28%. The higher accuracy of DenseNet121 suggests a better balance between precision and recall, leading to more correct predictions on the validation set.

InceptionV3 showed moderate performance with a validation accuracy of 56.58% and a validation loss of 0.3294. While it did not perform as well as DenseNet121 and Xception, it provides a valuable point of comparison due to its unique architectural features.

The transformer-based models showed varied results. The Swin Transformer achieved a validation accuracy of 60.51% with a validation loss of 0.2196, outperforming the Vision Transformer, which had a significantly higher validation loss (0.2746) and a low validation accuracy of 38.0% (*Figure 13*). The comparatively poor performance of ViT may be attributed to its higher data requirements and sensitivity to hyperparameter settings, which are critical factors in transformer architectures.

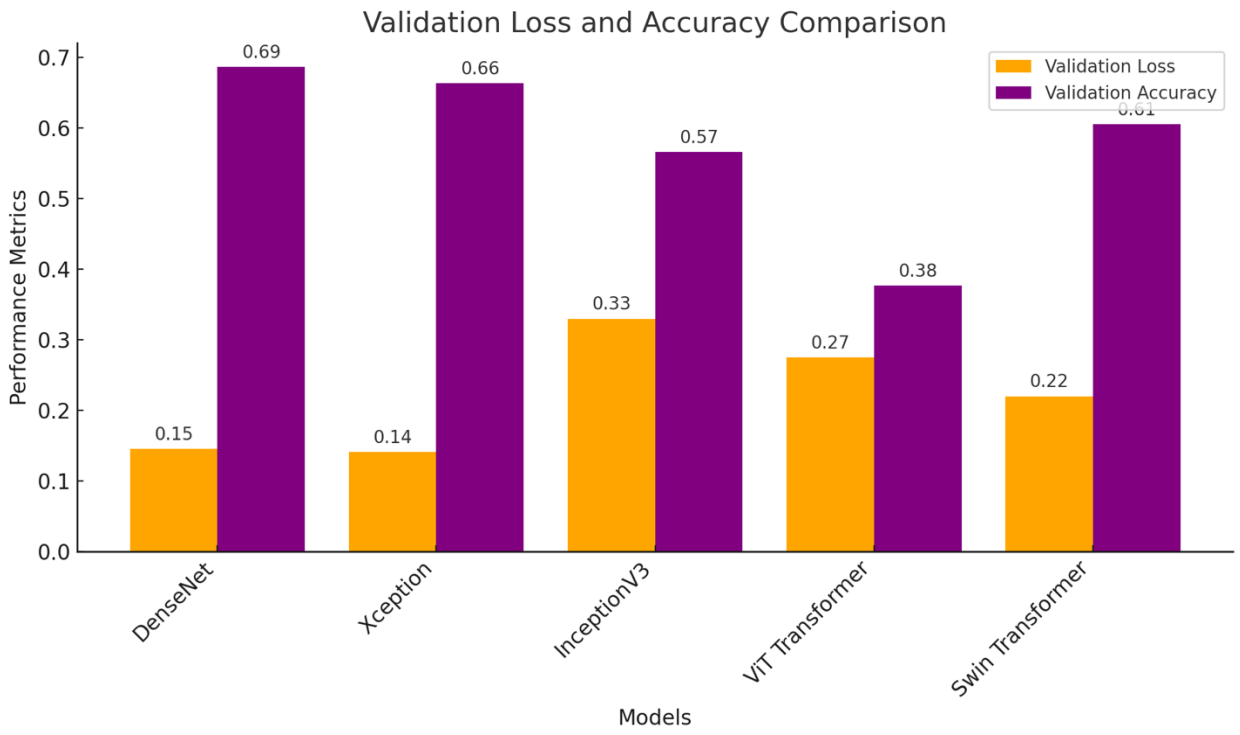


Figure 13 **: Validation loss and accuracy for various deep learning models**

Comparison of validation performance metrics across different deep learning architectures. Each pair of bars represents a single model, with the orange bar indicating the mean validation loss and the purple bar indicating the mean validation accuracy attained by that model. Lower validation loss and higher validation accuracy correspond to better predictive performance. Models shown from left to right are DenseNet, Xception, InceptionV3, ViT Transformer, and Swin Transformer. Numerical values above the bars specify the exact metric for each model, allowing direct performance comparisons

4.3 In-Depth Class-by-Class Performance Analysis

To gain deeper insights into the models' capabilities, we examined precision, recall, and F1-score for each subcellular localization class. Micro- and macro-averaged precision, recall, and F1-scores are shown for the five models evaluated on protein subcellular localization: Xception, InceptionV3, DenseNet, Swin Transformer, and ViT Transformer. Micro-averaging calculates metrics globally (summing over all classes), while macro-averaging treats each class equally (averaging per-class metrics). The data reflect each model's performance in correctly classifying proteins into multiple subcellular compartments (Table 3).

DenseNet121 demonstrated high F1-scores across most classes, indicating a balanced ability to correctly identify both major and minor subcellular locations (Figure 14). Notably, it achieved high precision and recall in well-represented classes such as Nucleoplasm (F1-score: 0.83), Cytosol (F1-score: 0.71), and Mitochondria (F1-score: 0.81). However, it showed relatively lower performance in classes like Nuclear bodies (F1-score: 0.59) and Intermediate filaments (F1-score: 0.48), suggesting challenges in capturing features for underrepresented or visually ambiguous classes.

Table 3: Micro and Macro Performance Metrics for Each Model.

Model	Precision (Micro)	Recall (Micro)	F1-Score (Micro)	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
Xception	0.82	0.67	0.74	0.81	0.57	0.65
InceptionV3	0.66	0.55	0.60	0.51	0.33	0.35
DenseNet	0.83	0.66	0.74	0.81	0.57	0.66
Swin	0.16	0.73	0.26	0.10	0.47	0.15
ViT	0.20	0.27	0.23	0.03	0.13	0.04

Xception performed comparably to DenseNet121 in major locations such as Nucleoplasm (F1-score: 0.83), Nucleoli (F1-score: 0.78), and Cytosol (F1-score: 0.75). It exhibited slightly better adaptability for some underrepresented classes, such as the Nucleoli fibrillar center (F1-score: 0.53), outperforming DenseNet121 in this category. However, Xception showed noticeable dips in performance for Nuclear bodies (F1-score: 0.44) and Nuclear speckles (F1-score: 0.71), indicating areas where it struggled with recall and could benefit from further optimization.

InceptionV3 underperformed compared to the other CNN models in several key classes. While it achieved acceptable performance in major classes like Nucleoplasm (F1-score: 0.79) and Cytosol (F1-score: 0.67), it showed significant deficiencies in classes such as nuclear membrane (F1-score: 0.00), Intermediate filaments (F1-score: 0.00), and Nucleoli fibrillar center (F1-score: 0.06). These

results suggest that InceptionV3 had difficulty handling minor or ambiguous subcellular locations effectively and may require substantial adjustments or further pretraining optimization to be viable for this task.

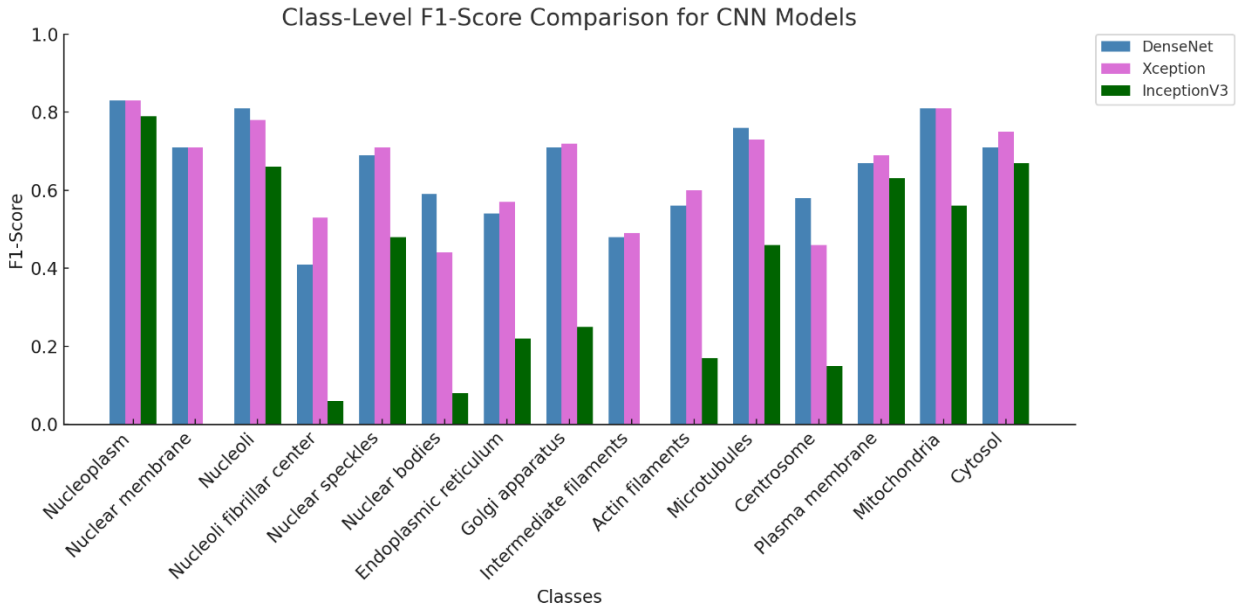


Figure 14: ***F1-score comparison for CNN-based architectures***

F1 Score Comparison. This bar chart compares the class-level F1-scores for three CNN-based architectures: DenseNet, Xception, and InceptionV3. Each bar along the x-axis group corresponds to one of the considered subcellular compartments. The F1-score, represented by the bar height, indicates the model's ability to correctly identify the class while minimizing false positives and false negatives. By examining these class-by-class results, it becomes possible to discern the strengths and weaknesses of each CNN architecture, as well as their relative capabilities in distinguishing specific cellular structures

The Swin Transformer exhibited moderate performance, with an overall validation accuracy of 60.51%. It demonstrated high recall but low precision in major classes like Nucleoplasm (precision: 0.38, recall: 0.99) and Cytosol (precision: 0.41, recall: 0.99), resulting in moderate F1-scores due to a high rate of false positives. This indicates that while the model could identify the most true instances of these classes, it incorrectly labeled many other instances as belonging to them (Figure 15).

The Vision Transformer performed poorly across all classes, with a validation accuracy of only 38.0%. It achieved high recall for Nucleoplasm (recall: 1.00) and Cytosol (recall: 1.00) but with very low precision (precision: 0.38 and 0.41, respectively), leading to moderate F1-scores due to

a high false-positive rate. The model failed to correctly identify instances for other classes, resulting in precision, recall, and F1-scores of 0.00 (Figure 15).

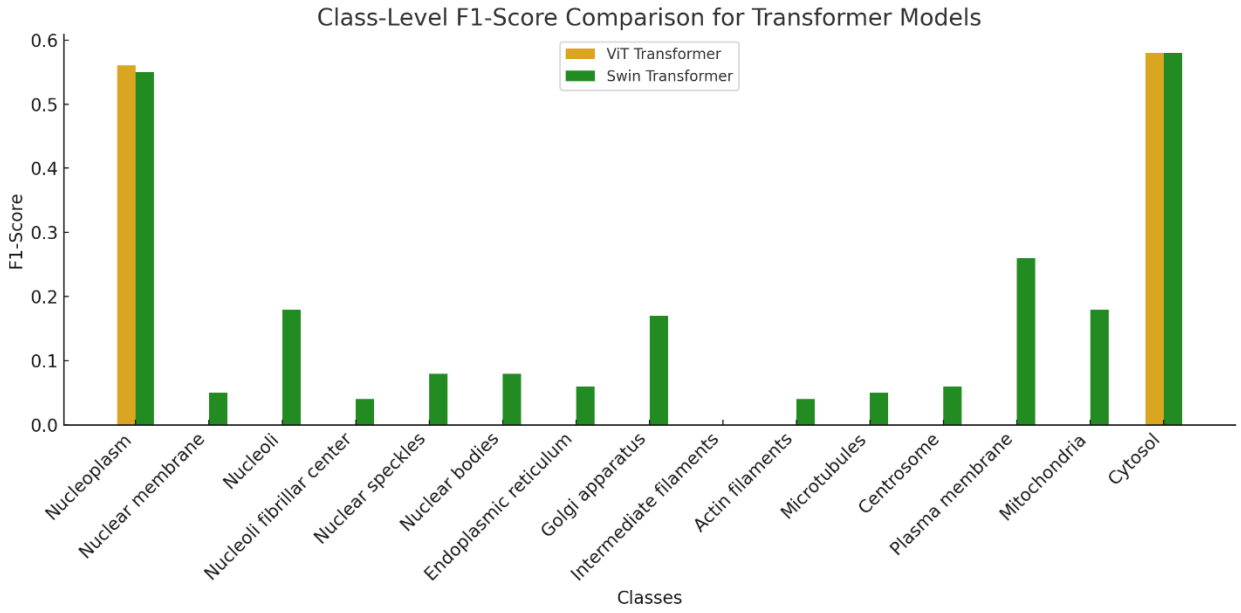


Figure 15 ***: F1-score comparison for transformer-based models***

F1 Score Comparison. This bar chart compares the class-level F1-scores for two transformer-based architectures: ViT Transformer and Swin Transformer. Each class along the x-axis represents a distinct subcellular compartment (e.g., Nucleoplasm, Nuclear membrane, Golgi apparatus), and the height of each bar indicates the F1-score achieved by that model on that specific class. By comparing the pairs of bars for each class, differences in the models' performance can be observed, highlighting which architectural design better captures the features of particular cellular structures

.

Confusion matrices are widely used tools for evaluating the performance of classification models, providing a summary of predictions made by a model compared to actual class labels. Each cell in the matrix represents the number of predictions for a specific class, highlighting true positives, false positives, and misclassifications (Provost & Kohavi, 1998). In this study, confusion matrices and classification metrics (precision, recall, and F1-score) were used to compare the performance of CNN-based models (DenseNet, InceptionV3, Xception) and Transformer-based models (Swin Transformer, ViT) in the task of subcellular localization (Figure 16).

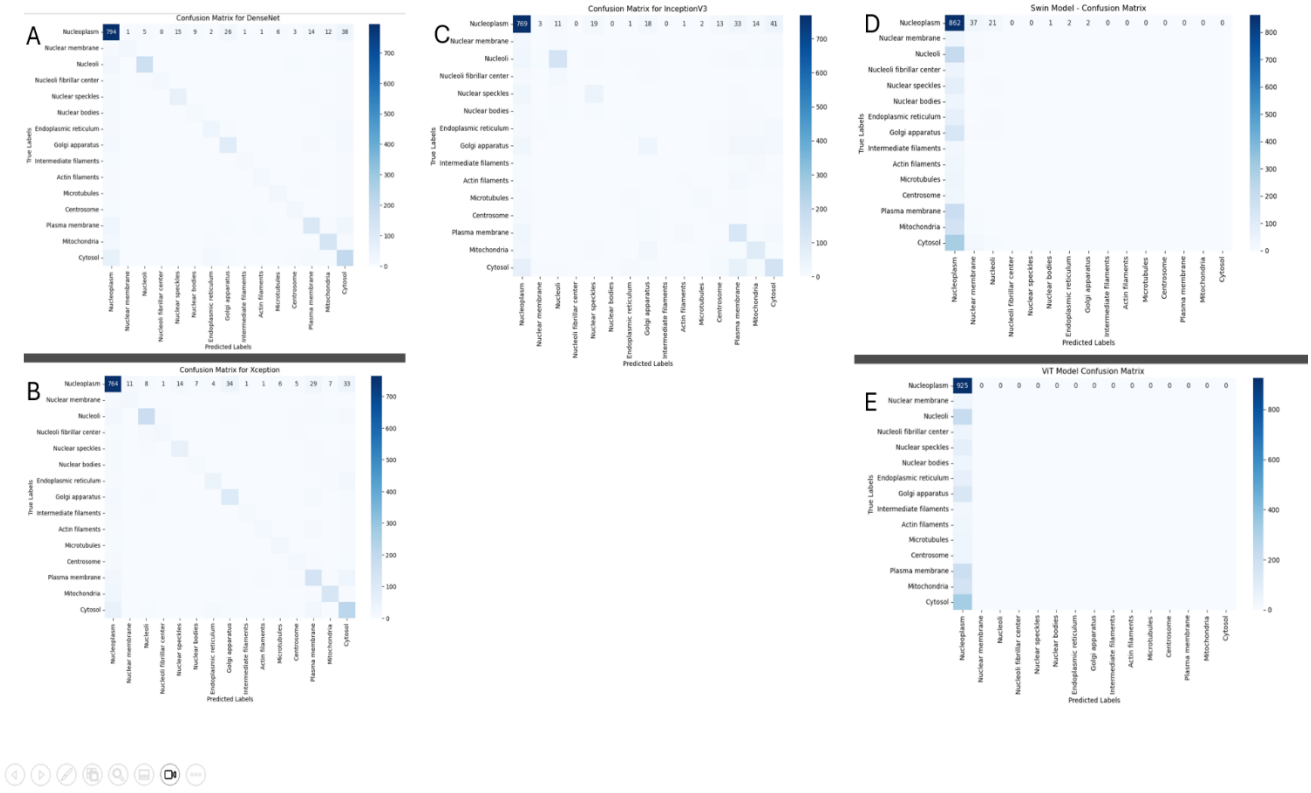


Figure 16 ***: Confusion matrices for five deep learning architectures***

Confusion matrix analysis. Confusion matrices showing classification performance of DenseNet,, Xception and InceptionV3, Swin Transformer and Vision Transformer (ViT) models for subcellular localization (A-E). The matrices illustrate the true labels (y-axis) against the predicted labels (x-axis) for different cellular compartments, with darker shades indicating higher counts

.

DenseNet emerged as the best-performing model among the CNN architectures, achieving a micro F1-score of 0.74 and a macro F1-score of 0.66. Its confusion matrix highlights strong classification accuracy for dominant classes, particularly "Nucleoplasm," with minimal misclassifications. However, DenseNet struggled with smaller and visually similar classes, such as "Nucleoli fibrillar center" and "Intermediate filaments," which had lower recall. Xception performed similarly to DenseNet, with a micro F1-score of 0.74 and a macro F1-score of 0.65, but its confusion matrix reveals more pronounced misclassifications for minority classes like "Nuclear bodies." InceptionV3, by contrast, showed the weakest performance among CNNs, with a micro F1-score of 0.60 and a macro F1-score of 0.35. Its confusion matrix illustrates widespread misclassification

across almost all minority classes, particularly "Nuclear membrane" and "Nucleoli fibrillar center," underscoring its difficulty in capturing subtle features within these categories.

Transformer-based models, Swin and ViT, underperformed relative to their CNN counterparts, with Swin achieving a micro F1-score of 0.26 and a macro F1-score of 0.15, while ViT reported a micro F1-score of 0.23 and a macro F1-score of 0.04. Swin showed moderate success in predicting "Nucleoplasm" and "Plasma membrane," as observed in its confusion matrix, but failed to generalize well across most minority classes. ViT exhibited severe overfitting to "Nucleoplasm," with nearly all other categories being misclassified into this dominant class. This limitation reflects the challenges transformer models face in handling highly imbalanced datasets without extensive fine-tuning.

Overall, the results reveal the significant impact of class imbalance on model performance. Dominant classes, such as "Nucleoplasm," received disproportionately higher precision and recall scores, as evidenced by the micro-average metrics in all models. In contrast, macro averages, which weigh all classes equally, highlighted the difficulty in accurately classifying underrepresented categories. CNN models, particularly DenseNet, excelled in leveraging localized feature extraction to distinguish between visually similar classes, whereas transformers, which rely on global attention mechanisms, struggled to capture the granularity required for subcellular localization tasks.

4.4 Precision-Recall Scatter Plot Evaluating Model Trade-Offs

We further evaluated the models using precision-recall scatter plots to gain deeper insights into how each model balanced precision and recall across different classes (Figure 17). In these plots, the dashed line represents a perfect balance where precision equals recall. Points closer to this line indicate a well-balanced trade-off, suggesting that the model accurately identifies true positives (high recall) while avoiding false positives (high precision).

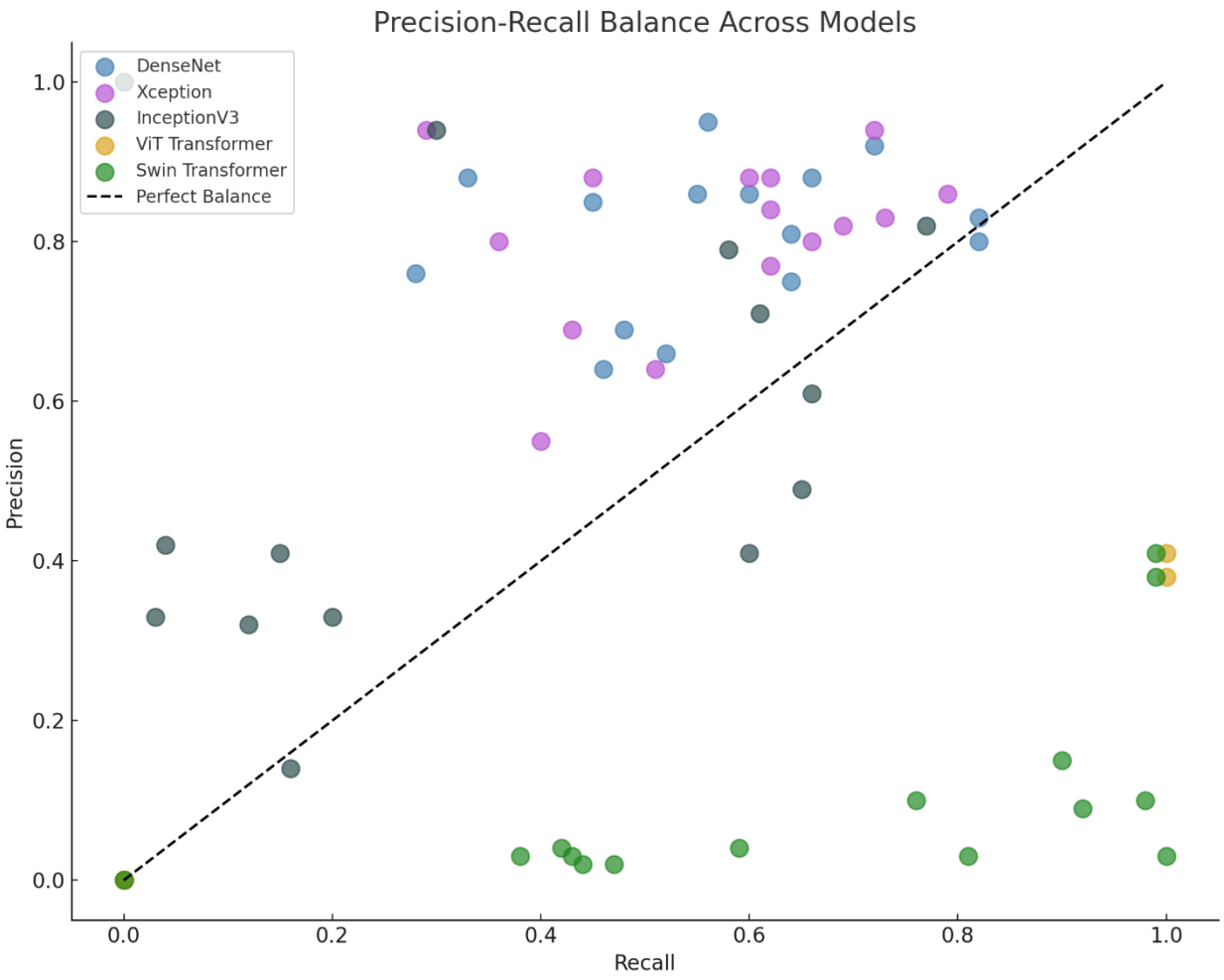


Figure 17: ***Precision-recall balance for various models and classes***

Precision-Recall Balance. This scatter plot displays the precision-recall balance of multiple models for various classes. Each point corresponds to a particular model's performance on a given class, with precision on the vertical axis and recall on the horizontal axis. The dashed line represents perfect balance (precision equals recall). Points above the line show where precision exceeds recall, while points below indicate recall outweighs precision. The different colors and markers represent various architectures (DenseNet, Xception, InceptionV3, ViT Transformer, and Swin Transformer), enabling a direct comparison of how each model trades off between capturing positive instances accurately (precision) and capturing as many positives as possible (recall).

DenseNet121 exhibited many points close to or above the balance line, especially for well-represented classes. This distribution indicates a good precision-recall balance across various classes, suggesting DenseNet121's reliability in identifying instances and minimizing false positives in subcellular localization. However, a few points were below the line, particularly in regions with lower precision and recall, indicating that DenseNet121 struggled with underrepresented or visually ambiguous classes.

Xception performed comparably to DenseNet121, with several points near the balance line. The model generally maintained a similar balance of precision and recall, demonstrating consistent performance across most classes. Notably, there were a few classes where Xception showed higher precision and lower recall, as indicated by points to the right of the line. This pattern suggests that Xception may slightly prioritize avoiding false positives over identifying every true positive in some instances, indicating a more conservative approach to specific class predictions.

InceptionV3 had most points falling far from the balance line, particularly in the low precision and recall region. Many of its points were significantly below the line, indicating difficulty in maintaining a balance between precision and recall. For certain classes, InceptionV3 exhibited either very low precision or very low recall—evident from points near the x or y-axis—suggesting that it struggled to identify these classes without a high misclassification rate accurately. This performance indicates that InceptionV3 may require substantial improvements or tuning to distinguish between certain subcellular locations reliably.

The Vision Transformer had points clustering near the recall axis at the bottom-right of the plot, showing very high recall but very low precision for some classes. This distribution indicates that the model frequently predicted positive instances but included many false positives. The imbalance between precision and recall highlights challenges in discriminating between classes, suggesting that ViT struggled to make accurate predictions despite identifying the most true positives.

Compared to ViT, the Swin Transformer achieved a better balance between precision and recall, with points distributed closer to the perfect balance line. While it still struggled with certain classes—evident from some points near the recall axis indicating high recall but low precision—the overall distribution suggests improved performance over ViT. The Swin Transformer's hierarchical architecture may have contributed to its enhanced ability to capture both local and global features, leading to a more balanced trade-off.

4.5 ROC curve analysis Assessing Discriminative Power

The provided set of ROC (Receiver Operating Characteristic) curves compares the performance of five different models: Xception, InceptionV3, DenseNet, Swin Transformer, and ViT Transformer, based on their ability to distinguish between classes. The Area Under the Curve

(AUC) metric is a summary statistic used to evaluate the overall model performance (Figure 18). Higher AUC values indicate better model performance, reflecting a higher true positive rate (sensitivity) for a lower false positive rate.

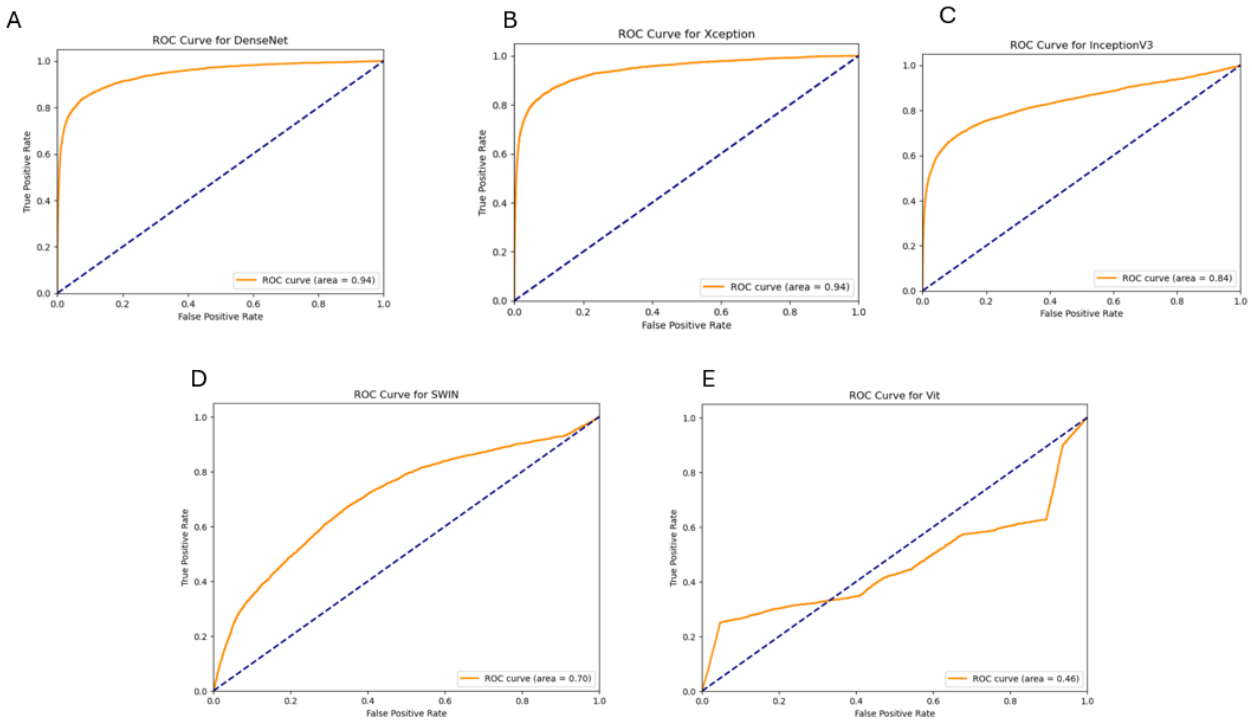


Figure 18 ***: ROC curves and AUC scores for each deep learning model***

ROC (Receiver Operating Characteristic) Area Under the Curve (AUC) metric is a summary. Each panel shows a Receiver Operating Characteristic (ROC) curve for a different deep learning model, including, DenseNet, Xception, InceptionV3, Swin Transformer, and ViT Transformer. The ROC curve plots the True Positive Rate (TPR) versus the False Positive Rate (FPR) across various decision thresholds. A curve closer to the top-left corner indicates better discriminative ability. The Area Under the Curve (AUC) value in each subplot provides a single numerical performance measure, with higher AUC values indicating superior overall classification accuracy.

The ROC curves highlight the superior performance of Xception and DenseNet, both achieving an AUC of 0.94, indicating excellent classification ability. InceptionV3 follows with an AUC of 0.84, demonstrating good performance but slightly less robust than the top two models. The Swin Transformer shows moderate capability with an AUC of 0.70, while the ViT Transformer performs poorly, with an AUC of 0.46, indicating that it is unsuitable for the current task. These results suggest that Xception and DenseNet are the most reliable models for the subcellular classification problem, with InceptionV3 as a viable alternative for further exploration.

4.6 Assessing Biological Relevance and Explainability of the Models using Grad-CAM

Deep neural networks often function as “black boxes,” making it non-trivial to confirm whether the network is leveraging biologically relevant features for its predictions. Visualization methods like Grad-CAM or attention rollouts help us inspect model decision-making by highlighting spatial regions most responsible for classification outcomes (Selvaraju et al., 2017b; Moujahid et al., 2021). If these highlighted regions correlate with bona fide subcellular structures (e.g., nucleoplasm, mitochondrial morphology), we gain confidence that the model is not just memorizing noise or artifacts. The top rows showcase the original immunofluorescence microscopy images for reference, representing subcellular structures such as the nucleoplasm, nuclear membrane, nucleoli, and more complex patterns like the Golgi apparatus, mitochondria, and cytosol. These images are essential benchmarks for evaluating the regions highlighted by the attention mechanism. The corresponding attention rollout/ GRAD-CAM heatmaps in the bottom rows reveal the regions of the input images the model attended to during classification (Figure 19-21).

Figure 19 presents Grad-CAM overlays for three convolutional neural network (CNN) architectures—DenseNet, Xception, and InceptionV3—applied to a diverse range of subcellular localization patterns. Each column corresponds to a specific class, with the top image displaying the original immunofluorescence micrograph and the subsequent rows showing Grad-CAM overlays from each model.

For classes that exhibit relatively straightforward morphological signatures (e.g., nucleoplasm, nucleoli, nuclear membrane), all three networks consistently localize their attention to biologically meaningful regions. The Grad-CAM hot spots (red/yellow regions) align well with the expected anatomical domains, suggesting that the networks rely on morphologically relevant features. This correspondence provides further evidence that the models are capturing and utilizing canonical patterns of subcellular architecture.

In contrast, more intricate classes such as Golgi apparatus, actin filaments, and centrosomes present greater interpretive challenges. Here, the Grad-CAM maps become more diffuse and

variable across architectures. DenseNet, Xception, and InceptionV3 may each emphasize slightly different subcellular regions, and the intensity of their attention may be less concentrated on well-defined morphological structures. This dispersion in the Grad-CAM signal suggests that these classes are more challenging for the models to delineate, possibly reflecting underlying biological complexity or subtle visual cues that differ significantly across samples.

Notably, differences across architectures also emerge. DenseNet's attention forms more coherent clusters, while Xception and InceptionV3 sometimes display more spatially distributed attention patterns. Such differences could stem from architectural design choices—such as dense connectivity in DenseNet or branching modules in InceptionV3—and their respective abilities to integrate multi-scale features.

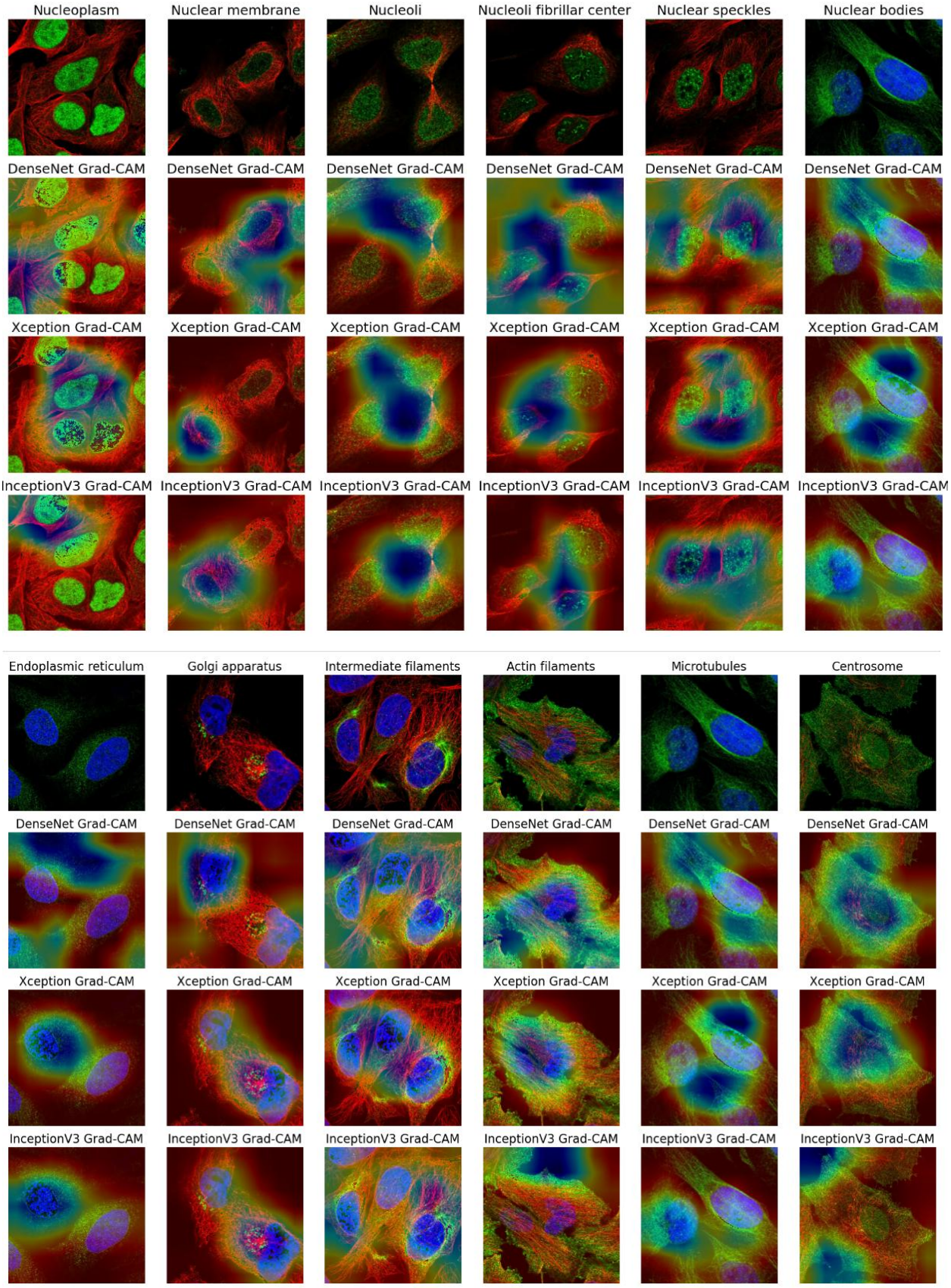

Nucleoplasm
Nuclear membrane
Nucleoli
Nucleoli fibrillar center
Nuclear speckles
Nuclear bodies
DenseNet Grad-CAM
Xception Grad-CAM
InceptionV3 Grad-CAM
Endoplasmic reticulum
Golgi apparatus
Intermediate filaments
Actin filaments
Microtubules
Centrosome

*Figure 19**: Grad-CAM visualizations for CNN models on subcellular structures***

Visualization methods with Grad-CAM or attention rollouts. The figure demonstrates Class Activation Maps (CAMs) generated using Grad-CAM for DenseNet, Xception, and InceptionV3 models. Each column represents a specific subcellular localization pattern, with the first row showcasing the original immunofluorescence microscopy images for reference. Easy patterns, such as nucleoplasm, nucleoli, and nuclear membrane, are visualized in the first set of columns, while more complex patterns, including Golgi apparatus, actin filaments, and centrosome, are shown in the latter columns. For each pattern, the Grad-CAM overlays highlight regions of interest used by the models for classification. Warmer colors (red/yellow) indicate regions with stronger model attention, while cooler colors (blue) represent lower attention. Models demonstrate consistent attention to biologically relevant regions in easy patterns but show variability in intricate patterns, reflecting differences in interpretability and performance.

Unlike CNNs, Vision Transformers and Swin Transformers employ self-attention instead of convolutional kernels. Classic Grad-CAM depends on the gradient flow through convolutional layers. Directly applying Grad-CAM to Transformers is less straightforward because attention—rather than localized receptive fields—underpins their feature representation. (Chefer, Gur, and Wolf, 2020; Chefer, Gur, and Wolf, 2021)

Swin Transformer refines ViT's global attention by restricting it to local windows. Each stage attends to features within small patch regions, making "pure attention rollout" trickier than in ViT. Instead, we can visualize per-window attention from the final stage (Figure 20). The heatmaps reveal discrete patches (red blobs) corresponding to how each window attends to local features. Since attention is computed within each local 7×7 patch region, the overlays appear "patchy" rather than continuous (Chefer, Gur, and Wolf, 2020; Cheng et al., 2021). Merging or stitching across multiple windows to form a single global map is non-trivial. Despite this, extracting pure attention still provides insight into which sub-regions each Swin window deems critical for classification (Cheng et al., 2021).

The class-level F1-score comparison (Figure 15) reinforces this interpretation. For classes like nucleoplasm and cytosol, where localized features are sufficient and relatively straightforward, the Swin model achieves competitive or higher F1-scores than ViT. This suggests that the localized attention mechanism helps effectively capture the salient features needed for classifying these simpler patterns.

However, for more complex or morphologically subtle classes (e.g., Golgi apparatus, intermediate filaments), the Swin model's performance still may not match the intuitive neatness of global attention maps. Instead, it scatters attention across multiple local patches, potentially indicating

difficulty consolidating these dispersed cues into a single, strong classification signal. Despite this, the Swin model often outperforms or matches ViT in F1-score for several challenging classes, implying that its windowed attention structure may help isolate and leverage fine-grained features otherwise lost in global averaging.

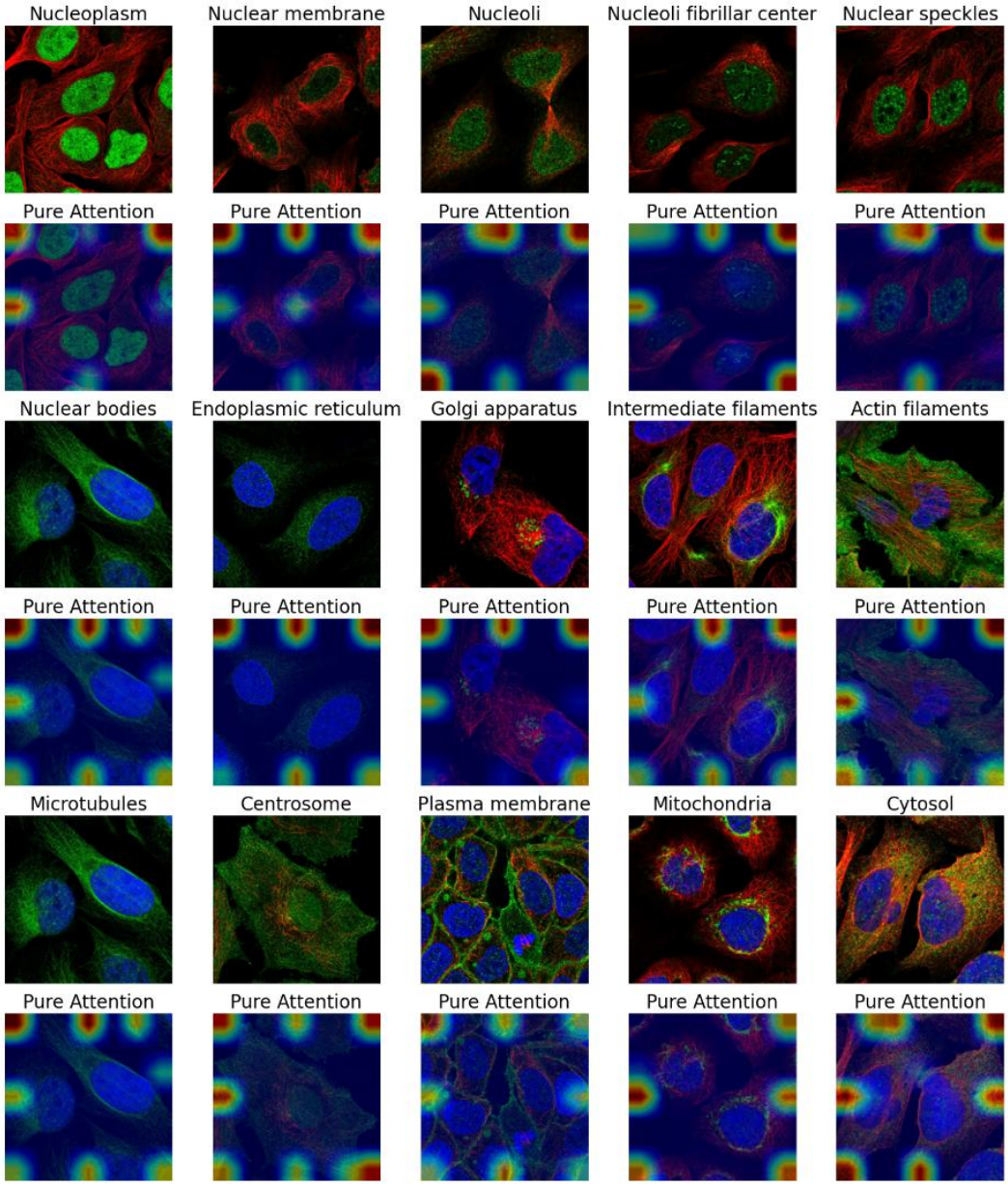


*Figure 20**: Swin Transformer attention maps across localization classes***

Visualization methods with Grad-CAM or attention rollouts. Visualization of local window attention in a Swin Transformer model for 15 subcellular localization classes. Top Rows: Original immunofluorescence microscopy images highlighting structures such as the nucleoplasm, nuclear membrane, nucleoli, and more complex compartments like the Golgi apparatus and mitochondria. Bottom Rows ("Pure Attention"): Overlayed attention maps extracted from the Swin Transformer's final stage. Because Swin uses local windows rather than global attention, each attention patch appears as distinct "blobs," illustrating the regions within each

window the model focuses on. Warmer areas (red/yellow) indicate higher local attention, suggesting more substantial relevance for classification. Cooler colors (blue) reflect lower attention. This "patchy" appearance reflects Swin's hierarchical, window-based architecture, as opposed to a single global map. Together, these images demonstrate how local self-attention mechanisms identify biologically meaningful features in each subcellular localization class

For ViT, attention is global: each patch can attend to all other patches in every self-attention layer. This structure permits attention rollout—an approach aggregating layer-by-layer attention weights to visualize how the model routes information from the class token to image patches. In the visualization of ViT attention rollout maps (Figure 21), we see how the Vision Transformer allocates its attention across various subcellular localization patterns. For more straightforward classes, such as nucleoplasm or cytosol, the rollout maps show the model consistently focusing on image regions that correspond to the expected biological structures. This alignment between attention hotspots and visually discernible cellular components suggests that ViT is leveraging relevant morphological cues for these more straightforward classes.

However, when faced with more complex patterns—such as the Golgi apparatus or intermediate filaments—the attention rollout maps become more diffuse or less intuitively aligned with specific features. This indicates that the model may struggle to pinpoint the key distinguishing patterns, potentially correlating with lower predictive performance for these classes.

The class-level F1-score comparison (Figure 15) provides quantitative support for the qualitative insights gleaned from the attention maps. For classes where ViT shows coherent and biologically meaningful attention patterns (e.g., nucleoplasm), the model also achieves relatively higher F1-scores. Conversely, for challenging categories like Golgi apparatus or less distinctive patterns, the F1-scores are noticeably lower, reflecting the difficulty in extracting consistent, discriminative features.

Compared to the Swin Transformer, the ViT model's performance exhibits some variance across classes. While ViT and Swin may both handle simpler patterns comparably well, Swin's windowed attention mechanism can sometimes better localize features in classes where ViT's global attention approach proves less effective. This disparity highlights that while ViT can excel at capturing global context, it may not always be optimal for classes requiring fine-grained spatial discrimination.

In summary, ViT's attention to rollout maps and class-level F1-scores reveal a pattern: the model excels in classes easily captured by global patch-to-patch relations but struggles in more intricate cases. These findings underscore the importance of selecting appropriate transformer architectures or feature extraction strategies based on the complexity of the biological structures under investigation.

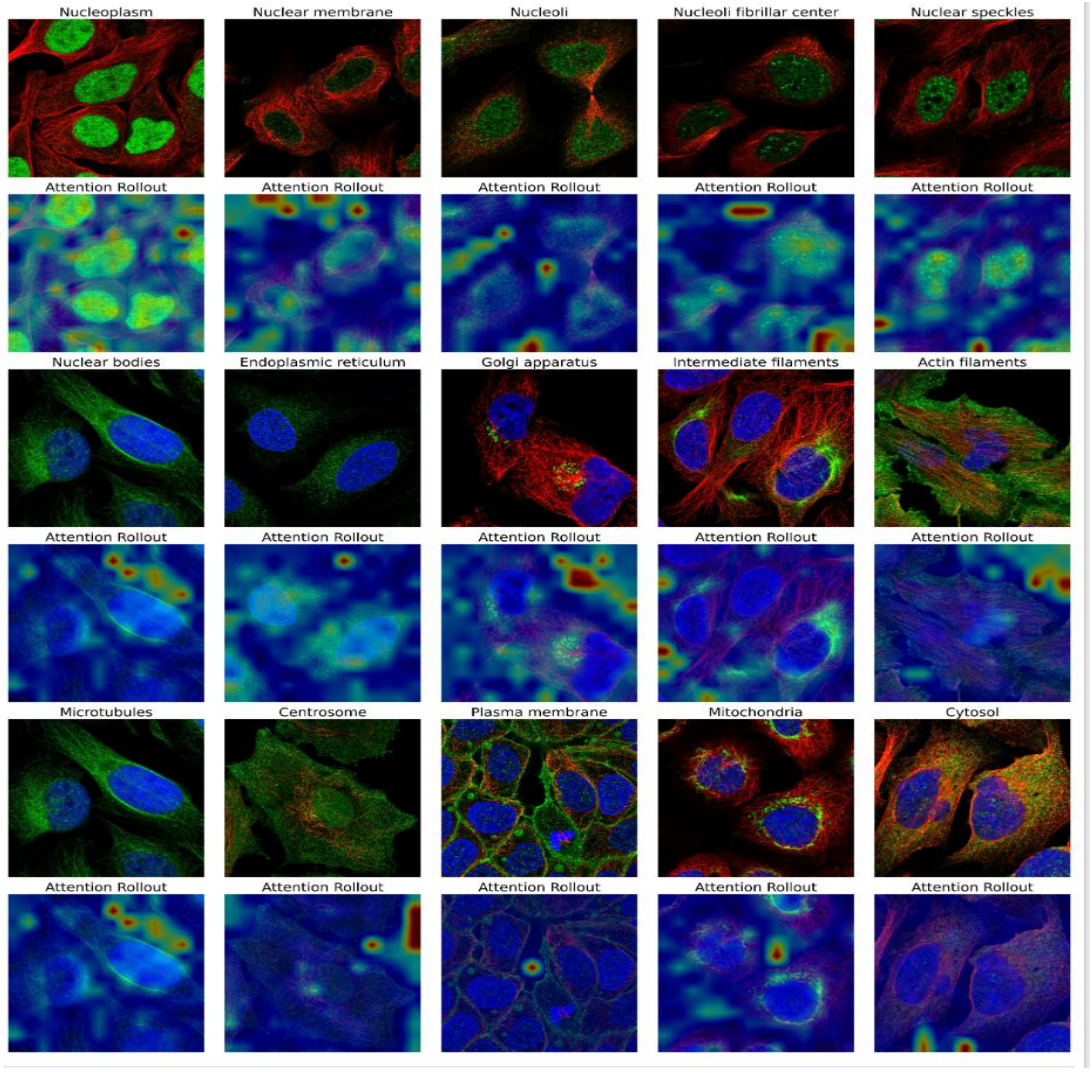


*Figure 21**: ViT attention rollout maps for different cellular compartments***

Visualization methods with Grad-CAM or attention rollouts. For classes like nucleoplasm or nucleoli, the rolled-out attention map aligns closely with the regions containing the relevant organelle features, reinforcing that ViT focuses on biologically valid signals. Categories such as mitochondria or Golgi apparatus show more diffuse or inconsistent attention distributions. The potential mismatch in localizing these structures is consistent with the model's lower accuracy on these classes and underscores the inherent difficulty of capturing morphological nuances in purely patch-based global attention. Warmer areas (red/yellow) indicate higher local attention, suggesting stronger relevance for classification. Cooler colors (blue) reflect lower attention

When comparing the Grad-CAM visualizations for the CNN models to the attention maps produced by the transformer-based models (ViT and Swin), distinct differences in interpretability emerge. The Grad-CAM overlays from DenseNet, Xception, and InceptionV3 typically form smooth, contiguous heatmaps that align closely with recognizable morphological features. In contrast, ViT's global attention maps and Swin's patch-based attention patterns often appear more fragmented, with key areas highlighted at the patch-level rather than forming a single coherent region. Although transformers may sometimes outperform CNNs in certain challenging classes, their attention maps reflect the more flexible, tokenized nature of their representations, making their visualization less straightforward to interpret. Nonetheless, each approach provides valuable insights, with CNN-based Grad-CAM offering more intuitive object-like saliency, while transformers reveal the underlying patch-level decision-making that can capture complex or subtle spatial cues.

Chapter Five -Discussion

5 Discussion

We trained and evaluated multiple deep learning architectures—spanning both convolutional neural networks (CNNs) and Transformers—to classify subcellular localization patterns in fluorescent microscopy images. Overall, the CNN-based approaches demonstrated superior performance, showing robust feature extraction capabilities for these complex tasks, whereas the Vision Transformer exhibited comparatively weaker results.

As summarized in Table 2, both DenseNet121 and Xception outperformed the other models in terms of validation loss and accuracy. Their superior performance can be linked to architectural features that facilitate efficient feature extraction (Banumathi et al., 2021; Albelwi, 2022). DenseNet121 employs dense connectivity, ensuring enhanced gradient flow and feature reuse, a crucial advantage for parsing complex fluorescent microscopy images (Huang et al., 2017; Hasan et al., 2021). Similarly, Xception leverages depthwise separable convolutions, allowing it to effectively learn spatial hierarchies with reduced computational overhead (Chollet, 2017).

These results indicate that DenseNet121 and Xception are particularly suited for multi-label subcellular protein localization tasks. DenseNet121's connectivity pattern alleviates the vanishing gradient problem and encourages learning nuanced details—an imperative factor when dealing with intricate fluorescence data. Meanwhile, Xception's separation of cross-channel and spatial correlations enables efficient feature extraction at multiple scales, offering another pathway to robust performance (Chollet, 2017; Lo, Yang, & Wang, 2019). The close similarity in validation loss for both models further suggests that each successfully minimizes error, supporting their applicability in this domain.

By contrast, InceptionV3 exhibited lower accuracy, possibly reflecting challenges in capturing underrepresented classes in a multi-label context. While Inception-style architectures have historically performed well (Szegedy et al., 2015; Nandini & Puviarasi 2021), Xception—derived

as an “Extreme Inception” variant—appears better optimized for this particular dataset through its novel use of depthwise separable layers (Wu, Liu, Yang, & Chen, 2020). This outcome underscores the importance of carefully matching architectural choices with task-specific demands, especially in specialized bioinformatics applications (Litjens et al., 2017a).

Turning to the transformer-based models, the Swin Transformer outperformed the Vision Transformer, yet still fell short of matching DenseNet121 and Xception’s validation accuracy (Table 2). The hierarchical architecture and shifted window mechanism of Swin Transformers likely enhanced its performance by capturing localized patterns more efficiently than Vision Transformers (Liu et al., 2021). This combination of efficiency and effectiveness makes Swin Transformers a notable improvement over Vision Transformers, especially in applications where scalability and computational resources are important factors (Liu et al., 2021; Feeser, n.d.). In contrast, Vision Transformers faced challenges with the dataset’s size and multi-label requirements, highlighting their dependence on large-scale datasets (Dosovitskiy et al., 2020). ViTs also demand considerable computational resources, particularly for high-resolution images, due to their attention mechanism, which compares all patches to one another (Kamsetty, Fricke & Liaw, 2020; Feeser, n.d.).

While transformers have demonstrated remarkable success in natural language processing and large-scale vision tasks (Vaswani et al., 2017), their application to specialized domains, such as subcellular protein localization, may require additional strategies to achieve optimal performance (Casola, Lauriola & Lavelli, 2022). For instance, transformers are known to be sensitive to hyperparameter choices and may necessitate extensive tuning to perform effectively in specialized tasks (Chen, 2022). Future work may explore larger training sets, more extensive hyperparameter searches (e.g., learning rate warm-up, weight decay schedules), or more aggressive data augmentations to unlock their potential fully (Chen, 2022). As a result, CNN-based models currently seem better aligned with the demands of multi-label subcellular protein localization (Table 2).

From a class-level perspective, DenseNet and Xception consistently deliver superior F1-scores, particularly on challenging compartments like nuclear structures and cytoskeletal filaments. InceptionV3 remains competitive in certain classes but shows noticeable dips in performance for

more intricate subcellular features. On the transformer side, Swin Transformer surpasses ViT for almost all classes, though both generally trail behind the best CNNs.

These observations underscore the importance of architectural choices and data availability when applying deep learning to specialized biomedical imaging tasks. CNN-based models—especially those with advanced connectivity or filter design—continue to be robust, while transformers, despite their success in other domains, may require further adjustments or larger datasets to unlock their potential in subcellular protein localization fully. However, because proteins often localize to multiple compartments (e.g., cytosol and nucleoplasm) simultaneously, the task requires multi-label classification rather than a single-label approach. Although we use binary cross-entropy and standard metrics like accuracy, F1-scores, and AUC, these can obscure crucial nuances of multi-label performance. For example, a model predicting only the most common label per sample might attain deceptively high accuracy in an imbalanced dataset. More specialized multi-label metrics—such as Hamming loss, subset accuracy, or average precision—can provide deeper insights, particularly for underrepresented compartments.

Grad-CAM and attention rollout analyses reveal that CNNs (DenseNet, Xception, InceptionV3) often generate smooth, contiguous activation maps aligning with known subcellular structures—an encouraging sign that these models rely on morphologically relevant features rather than spurious correlations. Meanwhile, transformer-based models exhibit more “patchy” or globally diffuse attention, reflecting their self-attention mechanisms. Though less visually intuitive, this token-by-token approach can sometimes detect patterns that CNN kernels miss. Confirming these attention maps with expert cell biologists would strengthen confidence that the learned features correspond to genuine biological patterns rather than artifacts

5.1 Biological Context and Impact

Accurate, high-throughput subcellular protein localization prediction has significant implications for basic biology and translational research. Mislocalized proteins are often implicated in diseases such as cancer and neurodegeneration (Wang & Wei, 2022; Barmada et al., 2010). Rapid computational screens can highlight candidate proteins for subsequent laboratory validation, saving time and expense compared to purely experimental methods (Xiao et al., 2024). As deep learning models become increasingly adept at parsing complex imaging data, they may guide

large-scale functional annotation of newly discovered proteins, accelerate drug discovery pipelines, and bolster personalized medicine approaches by revealing cell-type–specific localization phenomena (Jiang et al., 2021).

Overall, DenseNet121 and Xception stand out for their ability to capture complex spatial dependencies within microscopy images, while InceptionV3 and transformer-based models—particularly ViT—face limitations possibly tied to architecture-specific requirements and dataset size constraints. Although Swin Transformer showed moderate promise, further experimentation with data augmentation and hyperparameter tuning may be needed to close the performance gap. These findings align with previous studies highlighting the critical role of architectural design and training strategies when applying deep learning to specialized biomedical imaging tasks (Litjens et al., 2017b; Johnson & Khoshgoftaar 2019).

5.2 Limitations and Future Directions

Several constraints may have influenced model performance. First, although the Human Protein Atlas is a rich resource, it may not fully capture the diversity of cell types, imaging conditions, or staining protocols across different laboratories (Liimatainen et al., 2021). Second, transformers, especially ViT, often require massive datasets to converge well; while reasonably sized for biomedical standards, our dataset is still modest relative to ImageNet-scale corpora (Viso, 2024). Third, limited hyperparameter exploration—especially for ViT—may have capped performance. More fine-grained tuning (e.g., adjusting window sizes in Swin or altering the number of self-attention heads in ViT) could yield further improvements (Chandra et al., 2023).

Beyond hyperparameter optimization, combining multi-modal data (e.g., 3D z-stacks or time-lapse microscopy) could address subtle morphological features that 2D images may obscure (Alpert, 2023). Additionally, specialized data augmentations reflecting biological reality—such as morphological transformations or synthetic expansion of underrepresented classes—could help address the class imbalance and improve generalizability (Rana et al., 2023; Morales-Hernández, Van Nieuwenhuyse & Rojas Gonzalez, 2023).

In this study, We acknowledge that certain subcellular compartments are inherently underrepresent (Figure 10). However, we opted not to apply class imbalance remediation techniques (e.g.,

oversampling, undersampling, or class weighting) for several reasons. First, our goal was to preserve the intrinsic distribution of the dataset, which more accurately reflects real-world biological scenarios where some protein localizations are naturally rarer than others (Buda, Maki and Mazurowski 2018). Applying strong resampling or weighting methods can distort the true prevalence of each class and potentially inflate performance for classes that remain genuinely infrequent in practice (Wallace et al. 2011; Johnson and Khoshgoftaar 2019). Second, because our multi-label setup already combines overlapping compartments (e.g., cytosol and nucleoplasm), additional manipulation of class frequencies risked introducing further complexity into model training (He and Garcia 2009; Litjens et al. 2017a; Johnson and Khoshgoftaar 2019). Finally, the scale of the overall dataset—and the filtering step that removed classes with minimal samples—provided a sufficient baseline for each model to learn meaningful patterns without artificially balancing the classes (Johnson & Khoshgoftaar 2019). Nonetheless, future work could incorporate more targeted approaches (e.g., focal loss or cost-sensitive learning) if improving performance on rarely observed subcellular compartments becomes a primary objective (Chawla 2009).

Chapter Six- Conclusion

6 Conclusion

Our findings underscore that while transformer-based models continue to show promise—particularly with Swin's hierarchical attention—well-established CNNs like DenseNet121 and Xception currently lead in multi-label subcellular protein localization tasks under these experimental conditions. Optimizing architectures, dataset size, and hyperparameters could narrow the performance gap for transformers. In parallel, rigorous interpretability methods remain essential to ensure model decisions align with true biological markers and to further our understanding of disease mechanisms that arise from protein mislocalization. As image analysis and deep learning co-evolve, the synergy between computational predictions and experimental validation will prove invaluable for mapping the proteome's spatial organization within cells.

Future work should address class imbalance through advanced data augmentation strategies or oversampling techniques to improve recall for minority classes. Transformer models may benefit from hybrid architectures that integrate convolutional layers for finer spatial feature extraction while leveraging attention mechanisms for global context. Additionally, fine-tuning hyperparameters and applying regularization techniques may enhance the generalization ability of transformer models. These adjustments could further improve the robustness of both CNN and Transformer-based architectures in complex classification tasks.

7 Data availability

The dataset used for this analysis was obtained from the **Human Protein Atlas (HPA),** a publicly available repository of high-resolution immunofluorescence microscopy images. The dataset includes subcellular localization annotations for various proteins, facilitating large-scale deep learning-based predictions. Access to the dataset follows the terms and conditions set by the HPA for non-commercial research purposes.

8 Reference list

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M. & Ghemawat, S. (2016) 'TensorFlow: Large-scale machine learning on heterogeneous distributed systems', *arXiv:1603.04467 [cs.DC]* [Preprint]. Available at: https://arxiv.org/abs/1603.04467 (Accessed: 15 January 2025).
2. Abnar, S., & Zuidema, W. (2020). Quantifying attention flow in transformers. *ACL*.
3. Aerts, H.J., Velazquez, E.R., Leijenaar, R.T., et al., 2014. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5, p.4006.
4. Ahmed, M., Afreen, N., Ahmed, M., Sameer, M. and Ahamed, J., 2023. An inception V3 approach for malware classification using machine learning and transfer learning. *International Journal of Intelligent Networks*, 4, pp.11–18. https://doi.org/10.1016/j.ijin.2022.11.005.
5. Albawi, S., Mohammed, T.A. & Al-Zawi, S., 2017. Understanding of a convolutional neural network. *2017 International Conference on Engineering and Technology (ICET)*. DOI: https://doi.org/10.1109/ICEngTechnol.2017.8308186.
6. Albelwi, S.A., 2022. Deep Architecture based on DenseNet-121 Model for Weather Image Recognition. *International Journal of Advanced Computer Science and Applications*, 13(10), pp.559-563. Available at: <www.ijacsa.thesai.org> [Accessed 31 December 2024].
7. Alpert, C., 2023. Beyond hyperparameter optimization: Using AI to address digital implementation challenges. *Workshop on Machine Learning for CAD*. Available at: https://api.semanticscholar.org/CorpusID:264885586 [Accessed 9 January 2025].
8. Arulananth, T.S., Prakash, S.W., Ayyasamy, R.K., Kavitha, V.P., Kuppusamy, P.G. and Chinnasamy, P., 2024. Classification of Paediatric Pneumonia Using Modified DenseNet-121 Deep-Learning Model. *IEEE Access*, 12, pp.35716-35727. https://doi.org/10.1109/ACCESS.2024.3371151.
9. Banumathi, J., Muthumari, A., Dhanasekaran, S., Rajasekaran, S., Pustokhina, I.V., Pustokhin, D.A. and Shankar, K., 2021. An intelligent deep learning-based Xception model for hyperspectral image analysis and classification. Computers, Materials and Continua, 67(2), pp.2393-2407. Available at: https://doi.org/10.32604/cmc.2021.015605 [Accessed 31 Dec. 2024].
10. Barbe, L., Lundberg, E., Oksvold, P., Stenius, A., Lewin, E., Bjorling, E., et al. (2008) 'Toward a confocal subcellular atlas of the human proteome', *Molecular & Cellular Proteomics*, 7(3), pp. 499–508.

11. Barmada, S. J., Skibinski, G., Korb, E., Rao, E. J., Wu, J. Y., & Finkbeiner, S. (2010). Cytoplasmic Mislocalization of TDP-43 Is Toxic to Neurons and Enhanced by a Mutation Associated with Familial Amyotrophic Lateral Sclerosis. *Journal of Neuroscience, 30*(2), 639–649. https://doi.org/10.1523/JNEUROSCI.4988-09.2010
12. Bengio, Y., Simard, P. & Frasconi, P., 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5, pp.157–166. DOI: https://doi.org/10.1109/72.279181, PMID: 18267787.
13. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. & Linial, M., 2021. ProteinBERT: A Universal Deep-Learning Model of Protein Sequence and Function. *bioRxiv preprint*. DOI: https://doi.org/10.1101/2021.05.24.445464.
14. Brzozowski, R. S., White, M. L., & Eswara, P. J. (2020). Live-Cell Fluorescence Microscopy to Investigate Subcellular Protein Localization and Cell Morphology Changes in Bacteria. *Journal of Visualized Experiments, 153*, e59905. https://doi.org/10.3791/59905
15. Burns, T.J., Frei, A.P., Gherardini, P.F., Bava, F.A., Batchelder, J.E., Yoshiyasu, Y., et al. (2017) 'High-throughput precision measurement of subcellular localization in single cells', *Cytometry A*, 91(2), pp. 180–189.
16. Casola, S., Lauriola, I. & Lavelli, A., 2022. Pre-trained transformers: an empirical comparison. *Machine Learning with Applications*, 9, p.100334. DOI: 10.1016/j.mlwa.2022.100334.
17. Cham, B. (2023) Breaking Through the Darkness: How to Enhance Low-Light Images with Deep Learning Techniques. HTX S&S COE, 26 May. (If there is a URL, include it, for example: Available at: <URL> [Accessed 17 January 2025].)
18. Chandra, A., Tünnermann, L., Löfstedt, T. & Gratz, R., 2023. Transformer-based deep learning for predicting protein properties in the life sciences. *eLife*, 12, p.e82819. DOI: https://doi.org/10.7554/eLife.82819.
19. Chefer, H., Gur, S. and Wolf, L., 2020. Transformer Interpretability Beyond Attention Visualization. *arXiv preprint*. https://doi.org/10.48550/arXiv.2012.09838.
20. Chefer, H., Gur, S. and Wolf, L., 2021. Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers. *arXiv preprint*. Available at: https://arxiv.org/abs/2103.15679 [Accessed 9 January 2025].
21. Chen, H., Yang, Y. and Zhang, S., 2020. Learning Robust Scene Classification Model with Data Augmentation Based on Xception. *Journal of Physics: Conference Series*, 1575, p.012009. https://doi.org/10.1088/1742-6596/1575/1/012009.
22. Chen, Y., 2022. Optformer: Towards universal hyperparameter optimization with transformers. *Google Research Blog*, 18 August. Available at: https://research.google/blog/optformer-towards-

universal-hyperparameter-optimization-with-transformers/?utm_source=chatgpt.com [Accessed 9 January 2025].

23. Cheng, J., Bendjama, K., Rittner, K. & Malone, B., 2021. BERTMHC: improved MHC-peptide class II interaction prediction with transformer and multiple instance learning. *Bioinformatics*, 37, pp.4172–4179. DOI: https://doi.org/10.1093/bioinformatics/btab422, PMID: 34096999.
24. Chollet, F. (2015) 'Keras', GitHub. Available at: https://github.com/keras-team/keras (Accessed: 15 January 2025).
25. Chollet, F. (2017). Xception: Deep Learning with Depthwise Separable Convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1251–1258.
26. Chong, Y.T., Koh, J.L., Friesen, H., Duffy, S.K., Cox, M.J., Moses, A., et al. (2015) 'Yeast proteome dynamics from single cell imaging and automated analysis', *Cell*, 161(6), pp. 1413–1424.
27. Chowdhury R, Bouatta N, Biswas S, Floristean C, Kharkare A, Roye K. 2022. Single-sequenceprotein structure
28. Christoforou, A., Mulvey, C.M., Breckels, L.M., Geladaki, A., Hurrell, T., Hayward, P.C., et al. (2016) 'A draft map of the mouse pluripotent stem cell spatial proteome', *Nature Communications*, 7, p. 8992.
29. Clark, A. (2015). *Pillow (PIL Fork) Documentation*. https://pillow.readthedocs.io/en/stable/
30. Coursera Staff, 2024. *What Is Machine Learning? Definition, Types, and Examples*. Updated on 27 March 2024. Available at: <https://www.coursera.org/articles/what-is-machine-learning?msockid=3b2ef8adcd8462fe0221eb10cca76302> [Accessed 17 January 2025].
31. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V. & Le, Q.V., 2018. AutoAugment: learning augmentation policies from data. *Preprint* at https://arxiv.org/abs/1805.09501 [Accessed 2 October 2024].
32. Dai, Z., Yang, Z., Yang, Y., Cohen, W.W., Carbonell, J. & Le, Q.V., 2018. Transformer-XL: language modeling with longer-term dependency. *arXiv preprint*. Available at: https://arxiv.org/abs/1807.03819 [Accessed 2 October 2024].
33. Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J. & Kaiser, Ł., 2018. Universal Transformers. *arXiv preprint*. Available at: https://arxiv.org/abs/1807.03819 [Accessed 2 October 2024].
34. Devlin, J., Chang, M.W., Lee, K. & Toutanova, K., 2018. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint*. Available at: https://arxiv.org/abs/1810.04805 [Accessed 2 October 2024].
35. Dong, N., Zhao, L., Wu, C.H. and Chang, J.F., 2020. Inception v3 based cervical cell classification combined with artificially extracted features. *Applied Soft Computing*, 93, p.106311. Available at: https://doi.org/10.1016/j.asoc.2020.106311 [Accessed 31 Dec. 2024].

36. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint arXiv:2010.11929*.
37. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S., et al. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv*, arXiv:2010.11929.
38. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D. & Rost, B., 2020a. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning. *bioRxiv preprint*. DOI: https://doi.org/10.1101/2020.07.12.199554.
39. Falkner, S., Klein, A. & Hutter, F., 2018. BOHB: robust and efficient hyperparameter optimization at scale. in *35th International Conference on Machine Learning*, pp.1436–1445.
40. Feeser, S., n.d. Swin Transformer vs. Vision Transformer: Which One is Better for Vision Tasks?. *Stuart Feeser Blog*. Available at: https://stuartfeeser.com/blogs/ai-engineers/swin-vs-vit/index.html [Accessed 9 January 2025].
41. Feng, S., Sekine, S., Pessino, V., Li, H., Leonetti, M. D., & Huang, B. (2017). Improved Split Fluorescent Proteins for Endogenous Protein Labeling. *Nature Communications*, *8*, 370. https://doi.org/10.1038/s41467-017-00494-8
42. Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. NucleicAcids Research 39:W29–W37. DOI:https://doi.org/10.1093/nar/gkr367
43. Fukushima, K., 1980. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36, pp.193–202.
44. Gai, L., Xing, M., Chen, W., Zhang, Y. and Qiao, X. (2024) 'Comparing CNN-based and transformer-based models for identifying lung cancer: which is more effective?', *Multimedia Tools and Applications*, 83(20), pp. 59253–59269. Available at: https://doi.org/10.1007/s11042-023-17644-4.
45. Glorot, X., Bordes, A. & Bengio, Y., 2011. Deep sparse rectifier neural networks. In: *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, vol. 15, pp.315–323.
46. Godinez, W.J., Hossain, I., Lazic, S.E., Davies, J.W. & Zhang, X., 2017. A multi-scale convolutional neural network for phenotyping high-content cellular images. *Bioinformatics*, 33(13), pp.2010–2019.
47. Golubchik T, Wise MJ, Easteal S, Jermiin LS. 2007. Mind the gaps: evidence of bias in estimates of multiplesequence alignments. Molecular Biology and Evolution 24:2433–2442. DOI: https://doi.org/10.1093/molbev/msm176, PMID: 17709332

48. Gainza, P., Sverrisson, F., Monti, F., Rodolà, E., Boscaini, D., Bronstein, M.M., & Correia, B.E. (2020). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nature Methods, 17(2), 184-192. doi: 10.1038/s41592-019-0666-6.
49. Hanin, B., 2018. Which neural net architectures give rise to exploding and vanishing gradients? *Advances in Neural Information Processing Systems*.
50. Haralick, R. M., Shanmugam, K., & Dinstein, I. (1973). Textural Features for Image Classification. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-3*(6), 610–621. https://doi.org/10.1109/TSMC.1973.4309314
51. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2
52. Hasan, N., Bao, Y., Shawon, A. et al., 2021. DenseNet Convolutional Neural Networks Application for Predicting COVID-19 Using CT Image. *SN Computer Science*, 2, p.389. https://doi.org/10.1007/s42979-021-00782-7.
53. He, K., Zhang, X., Ren, S. & Sun, J., 2016. Deep residual learning for image recognition. in *IEEE Conference on Computer Vision and Pattern Recognition*, pp.770–778.
54. He, L., Zhang, S., Wu, L., Xia, H., Ju, F. & Zhang, H., 2021. Pre-Training Co-Evolutionary Protein Representation via A Pairwise Masked Language Model. *arXiv preprint*. Available at: https://arxiv.org/abs/2110.15527 [Accessed 2 October 2024].
55. Hofmarcher, M., Rumetshofer, E., Clevert, D.-A., Hochreiter, S. & Klambauer, G., 2019. Accurate prediction of biological assays with high-throughput microscopy images and convolutional networks. *Journal of Chemical Information and Modeling*, 59(3), pp.1163–1171.
56. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K.Q., 2017. Densely connected convolutional networks. in *IEEE Conference on Computer Vision and Pattern Recognition*, pp.4700–4708.
57. Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269.
58. Hung, V., Zou, P., Rhee, H.W., Udeshi, N.D., Cracan, V., Svinkina, T., et al. (2014) 'Proteomic mapping of the human mitochondrial intermembrane space in live cells via ratiometric APEX tagging', *Molecular Cell*, 55(2), pp. 332–341.
59. Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90-95.

60. Hutter, F., Kotthoff, L. & Vanschoren, J., 2019. Automated machine learning—methods, systems, challenges. Springer International Publishing.
61. Howard, J., & Gugger, S. (2020). *Deep Learning for Coders with fastai and PyTorch: AI Applications Without a PhD*. O'Reilly Media. (For general knowledge on CNNs and Transformers)
62. Ioffe, S. & Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. *Preprint* at https://arxiv.org/abs/1502.03167 [Accessed 2 October 2024].
63. Itzhak, D.N., Tyanova, S., Cox, J., Borner, G.H. (2016) 'Global, quantitative and dynamic mapping of protein subcellular localization', *Elife*, 5.
64. Jakobsen, L., Vanselow, K., Skogs, M., Toyoda, Y., Lundberg, E., Poser, I., et al. (2011) 'Novel asymmetrically localizing components of human centrosomes identified by complementary proteomics methods', *EMBO Journal*, 30(8), pp. 1520–1535.
65. Jiang, Y., Wang, D., Wang, W. & Xu, D., 2021. Computational methods for protein localization prediction. *Computational Structural Biotechnology Journal*, 19, pp.5834-5844. DOI: 10.1016/j.csbj.2021.10.023. PMID: 34765098; PMCID: PMC8564054.
66. Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on Deep Learning with Class Imbalance. *Journal of Big Data*, 6(1), 27.
67. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589. doi: 10.1038/s41586-021-03819-2
68. Kamsetty, A., Fricke, K. & Liaw, R., 2020. Hyperparameter optimization for transformers: A guide. *Medium*, Available at: https://medium.com/distributed-computing-with-ray/hyperparameter-optimization-for-transformers-a-guide-c4e32c6c989b [Accessed 9 January 2025].
69. Kingma, D.P. and Ba, J. (2015) 'Adam: A Method for Stochastic Optimization', *International Conference on Learning Representations (ICLR)*. arXiv:1412.6980 [Preprint]. Available at: https://arxiv.org/abs/1412.6980
70. Kohnhorst, C. L., Schmitt, D. L., Sundaram, A., & An, S. (2016). Subcellular Functions of Proteins under Fluorescence Single-Cell Microscopy. *Biochimica et Biophysica Acta (BBA)—Proteins and Proteomics*, *1864*(1), 77–84. https://doi.org/10.1016/j.bbapap.2015.10.001

71. Koumakis, L., 2020. Deep learning models in genomics; are we there yet? *Computational and Structural Biotechnology Journal*, 18, pp.1466–1473. DOI: https://doi.org/10.1016/j.csbj.2020.06.017 [Accessed 2 October 2024].
72. Lee, S.Y., Kang, M.G., Park, J.S., Lee, G., Ting, A.Y., Rhee, H.W. (2016) 'APEX fingerprinting reveals the subcellular localization of proteins of interest', *Cell Reports*, 15(8), pp. 1837–1847.
73. Li, J., Xiong, L., Schneider, J., & Murphy, R. F. (2012). Protein Subcellular Location Pattern Classification in Cellular Images Using Latent Discriminative Models. *Bioinformatics, 28*(12), i32–i39. https://doi.org/10.1093/bioinformatics/bts218
74. Lin, T.Y., Goyal, P., Girshick, R., He, K. & Dollár, P., 2017. Focal loss for dense object detection. in *IEEE International Conference on Computer Vision*, pp.2980–2988.
75. Linkon, A.H.M., Labib, M.M., Hasan, T., Hossain, M. and Jannat, M.-E., 2021. Deep learning in prostate cancer diagnosis and Gleason grading in histopathology images: An extensive study. *Informatics in Medicine Unlocked*, 24, p.100582. https://doi.org/10.1016/j.imu.2021.100582.
76. Litjens, G., Kooi, T., Bejnordi, B. E., et al. (2017). A Survey on Deep Learning in Medical Image Analysis. *Medical Image Analysis*, 42, 60–88.
77. Liu, G.-H., Zhang, B.-W., Qian, G., Wang, B., Mao, B., & Bichindaritz, I. (2020). Bioimage-Based Prediction of Protein Subcellular Location in Human Tissue with Ensemble Features and Deep Networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 17*(6), 1966–1980. https://doi.org/10.1109/TCBB.2019.2919961
78. Liu, T., Lin, Y., Wen, X., Jorissen, R.N. & Gilson, M.K., 2007. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Research*, 35, pp.D198–D201. DOI: https://doi.org/10.1093/nar/gkl999 [Accessed 2 October 2024].
79. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. & Guo, B. (2021) 'Swin Transformer: Hierarchical vision transformer using shifted windows', Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, 11–17 October, pp. 10012–10022.
80. Liu, Z., Qian, S., Xia, C., & Wang, C. (2024). Are transformer-based models more robust than CNN-based models? *Neural Networks, 172*, 106091. https://doi.org/10.1016/j.neunet.2023.12.045
81. Lo, W.W., Yang, X. and Wang, Y., 2019. An Xception convolutional neural network for malware classification with transfer learning. *NTMS 2019 - 10th IFIP International Conference on New Technologies, Mobility and Security*, pp.1-5. Available at: https://doi.org/10.1109/NTMS.2019.8763852 [Accessed 31 Dec. 2024].

82. Long, W., Yang, Y., & Shen, H.-B. (2020). ImPLoc: A Multi-Instance Deep Learning Model for the Prediction of Protein Subcellular Localization Based on Immunohistochemistry Images. *Bioinformatics, 36*(8), 2244–2250. https://doi.org/10.1093/bioinformatics/btz864
83. Lundberg, E., & Borner, G. H. H. (2019). Spatial Proteomics: A Powerful Discovery Tool for Cell Biology. *Nature Reviews Molecular Cell Biology, 20*(4), 285–302. https://doi.org/10.1038/s41580-019-0108-1
84. McKinsey & Company, 2024. *What is machine learning?*. 30 April 2024. Available at: <https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-machine-learning> [Accessed 17 January 2025].
85. Mikolov, T., Chen, K. & Corrado, G.S., 2013a. Efficient Estimation of Word Representations in Vector Space. *International Conference on Learning Representations*.
86. Mikolov, T., Kombrink, S., Burget, L. & Černocký, J., 2011b. Extensions of recurrent neural network language model. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. DOI: https://doi.org/10.1109/ICASSP.2011.5947611 [Accessed 2 October 2024].
87. Moen, E., Bannon, D., Kudo, T., Graf, W., Covert, M. & Van Valen, D., 2019. Deep learning for cellular image analysis. *Nature Methods*, [online] Available at: https://doi.org/10.1038/s41592-019-0403-1 [Accessed 2 October 2024].
88. Morales-Hernández, A., Van Nieuwenhuyse, I. & Rojas Gonzalez, S., 2023. A survey on multi-objective hyperparameter optimization algorithms for machine learning. *Artificial Intelligence Review*, 56, pp.8043–8093. DOI: 10.1007/s10462-022-10359-2.
89. Moujahid, H., Cherradi, B., Al-Sarem, M., Bahatti, L., Bakr, A., Assedik, M., Eljialy, A., Alsaeedi, A. and Saeed, F., 2021. Combining CNN and Grad-Cam for COVID-19 Disease Prediction and Visual Explanation. *Intelligent Automation and Soft Computing*, 32, pp.723-745.
90. Nair, V. & Hinton, G.E., 2010. Rectified linear units improve restricted Boltzmann machines. In: *Proceedings of the 27th International Conference on Machine Learning*. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.165.6419&rep=rep1&type=pdf [Accessed 23 January 2018].
91. Nambiar, A., Heflin, M., Liu, S., Maslov, S., Hopkins, M. & Ritz, A., 2020. Transforming the Language of Life: Transformer Neural Networks for Protein Prediction Tasks. In: *BCB '20: Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp.1–8. DOI: https://doi.org/10.1145/3388440.3412467.

92. Nandini, B. and Puviarasi, R., 2021. Detection of skin cancer using Inception V3 and Inception V4 convolutional neural network (CNN) for accuracy improvement. *Revista Gestão Inovação e Tecnologias*, 11(4).
93. Nanni, L., Lumini, A., & Brahnam, S. (2012). Survey on LBP Based Texture Descriptors for Image Classification. *Expert Systems with Applications*, *39*(3), 3634–3641. https://doi.org/10.1016/j.eswa.2011.09.030
94. Newberg, J., & Murphy, R. F. (2008). A Framework for the Automated Analysis of Subcellular Patterns in Human Protein Atlas Images. *Journal of Proteome Research, 7*(6), 2300–2308. https://doi.org/10.1021/pr7008117
95. Orre, L.M., Vesterlund, M., Pan, Y., Arslan, T., Zhu, Y., Fernandez Woodbridge, A., et al. (2019) 'Proteome-wide mapping of protein localization and relocalization', *Molecular Cell*, 73(1), pp. 166–182 e167.
96. Ouyang, W., Winsnes, C. F., Hjelmare, M., Cesnik, A. J., Åkesson, L., Xu, H., Sullivan, D. P., Dai, S., Lan, J., Jinmo, P., et al. (2019). Analysis of the Human Protein Atlas Image Classification Competition. *Nature Methods*, *16*(12), 1254–1261. https://doi.org/10.1038/s41592-019-0658-6
97. Pärnamaa, T., & Parts, L. (2017). Accurate Classification of Protein Subcellular Localization from High-Throughput Microscopy Images Using Deep Learning. *G3: Genes, Genomes, Genetics, 7*(5), 1385–1392. https://doi.org/10.1534/g3.117.039008
98. Pascanu, R., Mikolov, T. & Bengio, Y., 2013. On the difficulty of training recurrent neural networks. In: *International Conference on Machine Learning*. PMLR.
99. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N. & Antiga, L., 2017. Automatic differentiation in PyTorch. in *NIPS 2017 Autodiff Workshop*.
100. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. & Zettlemoyer, L., 2018. Deep Contextualized Word Representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. DOI: https://doi.org/10.18653/v1/N18-1202.
101. Phuong TM, Do CB, Edgar RC, Batzoglou S. 2006. Multiple alignment of protein sequences with repeats and rearrangements. Nucleic Acids Research 34:5932–5942. DOI: https://doi.org/10.1093/nar/gkl511, PMID:17068081
102. Provost, F., & Kohavi, R. (1998). Glossary of terms: Machine learning. *Journal of Machine Learning*, 30(2–3), 271–274. https://doi.org/10.1023/A:1007465528199
103. Pratap, A. (2023, January 16). Comparing CNNs and Transformers: Understanding the Differences and Key Components of These Popular Deep Learning Architectures. *Deep Learners*

in Deep Learning and Machine Learning on Medium. Retrieved from https://medium.com/deep-learners-in-deep-learning-and-machine/comparing-cnns-and-transformers-understanding-the-differences-and-key-components-of-these-popular-4cecdec2d0d9

104. prediction using a language model and deep learning. Nature Biotechnology 22:1–7. DOI: https://doi.org/10.1038/s41587-022-01432-w

105. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C. & Dosovitskiy, A., 2021. Do vision transformers see like convolutional neural networks. *Advances in Neural Information Processing Systems*.

106. Rahadianti, L., Azizah, A.Y. and Deborah, H. (2021) Evaluation of the quality indicators in dehazed images: Color, contrast, naturalness, and visual pleasingness. Heliyon, 7(9), p.e08038. https://doi.org/10.1016/j.heliyon.2021.e08038

107. Ramachandran, P., Zoph, B. & Le, Q.V., 2017. Searching for activation functions. *arXiv preprint*. Available at: https://arxiv.org/pdf/1710.05941.pdf [Accessed 23 January 2018].

108. Rana, P., Sowmya, A., Meijering, E. & Song, Y., 2023. Imbalanced classification for protein subcellular localization with multilabel oversampling. *Bioinformatics*, 39(1), p.btac841. DOI: 10.1093/bioinformatics/btac841. PMID: 36579866; PMCID: PMC9825308.

109. Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P. & Song, Y.S., 2019. Evaluating protein transfer learning with TAPE. *Advances in Neural Information Processing Systems*, pp.9689–9701.

110. Rasheed, M.T., Shi, D. and Khan, H. (2023) A comprehensive experiment-based review of low-light image enhancement methods and benchmarking low-light image quality assessment. Signal Processing, 204, p.108821. https://doi.org/10.1016/j.sigpro.2022.108821

111. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J. & Fergus, R., 2021. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118, p.e2016239118. DOI: https://doi.org/10.1073/pnas.2016239118, PMID: 33876751.

112. Roell, J. (n.d.). The Ultimate Guide: RNNs vs. Transformers vs. Diffusion Models. *Medium*. Retrieved from https://medium.com/@roelljr/the-ultimate-guide-rnns-vs-transformers-vs-diffusion-models-5e841a8184f3

113. Rustamy, F. (2023, June 4). Vision Transformers vs. Convolutional Neural Networks. *Medium*. Retrieved from https://medium.com/@faheemrustamy/vision-transformers-vs-convolutional-neural-networks-5fe8f9e18efc

114. Saethang T, Payne DM, Avihingsanon Y, Pisitkun T. 2016. A machine learning strategy for predicting localizationof post-translationalmodification sites in protein-protein interacting regions. BMC Bioinformatics 17:307. DOI:https://doi.org/10.1186/s12859-016-1165-8, PMID: 27534850

115. Schmiedel JM, Lehner B. 2019. Determining protein structures using deep mutagenesis. Nature Genetics51:1177–1186. DOI: https://doi.org/10.1038/s41588-019-0431-x, PMID: 31209395

116. Selvaraju, R. R., et al. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. ICCV, 618-626.

117. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D. and Batra, D., 2017b. Grad-CAM: Why did you say that? *arXiv preprint*. Available at: https://arxiv.org/abs/1611.07450 [Accessed 9 January 2025].

118. Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011). On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases* (ECML PKDD 2011) (pp. 145–158). Springer.

119. Sharma A, Lyons J, Dehzangi A, Paliwal KK. 2013. A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. Journal of Theoretical Biology 320:41–46. DOI:https://doi.org/10.1016/j.jtbi.2012.12.008, PMID: 23246717

120. Sharma, S. and Kumar, S., 2022. The Xception model: A potential feature extractor in breast cancer histology images classification. *ICT Express*, 8(1), pp.101-108. Available at: https://doi.org/10.1016/j.icte.2021.11.010 [Accessed 31 Dec. 2024].

121. Smith, L.N., 2017. Cyclical learning rates for training neural networks. in *IEEE Winter Conference on Applications of Computer Vision*, pp.464–472.

122. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15, pp.1929–1958.

123. Stadler, C., Skogs, M., Brismar, H., Uhlen, M., Lundberg, E. (2010) 'A single fixation protocol for proteome-wide immunofluorescence localization studies', *Journal of Proteomics*, 73(6), pp. 1067–1078.

124. Stewart, M. (2007). Molecular Mechanism of the Nuclear Protein Import Cycle. *Nature Reviews Molecular Cell Biology, 8*(3), 195–208. https://doi.org/10.1038/nrm2114

125. Su, R., He, L., Liu, T., Liu, X., & Wei, L. (2021). Protein Subcellular Localization Based on Deep Image Features and Criterion Learning Strategy. *Briefings in Bioinformatics, 22*(4), bbaa313. https://doi.org/10.1093/bib/bbaa313

126. Szegedy, C., Liu, W., Jia, Y., et al. (2015). Going Deeper with Convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9.

127. Tahir, M., Khan, A., & Majid, A. (2012). Protein Subcellular Localization of Fluorescence Imagery Using Spatial and Transform Domain Features. *Bioinformatics*, *28*(1), 91–97. https://doi.org/10.1093/bioinformatics/btr618

128. The Human Protein Atlas, n.d. *The Human Protein Atlas*. Available at: https://www.proteinatlas.org/ [Accessed 17 January 2025].

129. Thul, P. J., & Lindskog, C. (2018). The Human Protein Atlas: A Spatial Map of the Human Proteome. *Protein Science*, *27*(1), 233–244. https://doi.org/10.1002/pro.3307

130. Thul, P.J., Akesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., et al. (2017) 'A subcellular map of the human proteome', *Science*, 356(6340).

131. Trent. (2018). *iterative stratification [Software]*. GitHub. https://github.com/trent-b/iterative-stratification

132. Ullah, M., Han, K., Hadi, F., Xu, J., Song, J., & Yu, D.-J. (2021). PScL-HDeep: Image-Based Prediction of Protein Subcellular Location in Human Tissue Using Ensemble Learning of Handcrafted and Deep Learned Features with Two-Layer Feature Selection. *Briefings in Bioinformatics, 22*(5), bbab278. https://doi.org/10.1093/bib/bbab278

133. UniProt. 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research 49:D480–D489. DOI: https://doi.org/10.1093/nar/gkaa1100

134. Vanschoren, J., 2018. Meta-learning: a survey. *Preprint* at https://arxiv.org/abs/1810.03548 [Accessed 2 October 2024].

135. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems (NeurIPS)*, 5998–6008.

136. Väth, P., Münch, M., Raab, C. & Schleif, F.M., 2022. PROVAL: A framework for comparison of protein sequence embeddings. *Journal of Computational Mathematics and Data Science*, 2022, p.100044. DOI: https://doi.org/10.1016/j.jcmds.2022.100044.

137. Viso, 2024. *Xception model: An introduction to deep learning architecture*. Available at: https://viso.ai/deep-learning/xception-model/ [Accessed 31 Dec. 2024].

138. Wang, F., & Wei, L. (2022). Multi-Scale Deep Learning for the Imbalanced Multi-Label Protein Subcellular Localization Prediction Based on Immunohistochemistry Images. *Bioinformatics, 38*(10), 2602–2611. https://doi.org/10.1093/bioinformatics/btac065

139. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, C., Rault, T., Louf, R., Funtowicz, M. & Brew, J. (2020) 'Transformers: State-of-the-art natural language

processing', *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, 16–20 November, pp. 38–45.

140. Wu, X., Liu, R., Yang, H. and Chen, Z., 2020. An Xception-based convolutional neural network for scene image classification with transfer learning. *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*, Guangzhou, China, pp.262-267. Available at: https://doi.org/10.1109/ITCA52113.2020.00063 [Accessed 31 Dec. 2024].

141. Xiang, L., Yang, Q.-L., Xie, B.-T., Zeng, H.-Y., Ding, L.-J., Rao, F.-Q., Yan, T., Lu, F., Chen, Q., & Huang, X.-F. (2023). Dysregulated Arginine Metabolism Is Linked to Retinal Degeneration in Cep250 Knockout Mice. *Investigative Ophthalmology & Visual Science, 64*(2), 2. https://doi.org/10.1167/iovs.64.2.2

142. Xu, Y.-Y., Yao, L.-X., & Shen, H.-B. (2018). Bioimage-Based Protein Subcellular Location Prediction: A Comprehensive Review. *Frontiers of Computer Science*, *12*(1), 26–39. https://doi.org/10.1007/s11704-016-6035-2

143. Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. Insights Imaging. 2018 Aug;9(4):611-629. doi: 10.1007/s13244-018-0639-9. Epub 2018 Jun 22. PMID: 29934920; PMCID: PMC6108980.

144. Zeng, H., Edwards, M.D., Liu, G. & Gifford, D.K., 2016. Convolutional neural network architectures for predicting DNA-protein binding. *Bioinformatics*, 32(Suppl 1), pp.i121–i127. DOI: https://doi.org/10.1093/bioinformatics/btw255 [Accessed 2 October 2024].

145. Zhao, C., Xu, Z., Wang, X., Tao, S., MacDonald, W. A., He, K., Poholek, A. C., Chen, K., Huang, H., & Chen, W. (2024). Innovative Super-Resolution in Spatial Transcriptomics: A Transformer Model Exploiting Histology Images and Spatial Gene Expression. *Briefings in Bioinformatics, 25*(1), bbae052. https://doi.org/10.1093/bib/bbae052

146. Ziff, O. J., Harley, J., Wang, Y., Neeves, J., Tyzack, G., Ibrahim, F., Skehel, M., Chakrabarti, A. M., Kelly, G., & Patani, R. (2023). Nucleocytoplasmic mRNA Redistribution Accompanies RNA Binding Protein Mislocalization in ALS Motor Neurons and Is Restored by VCP ATPase Inhibition. *Neuron, 111*(15), 3011–3027.e7. https://doi.org/10.1016/j.neuron.2023.04.011

147. Zhang, Z., Xu, M., Jamasb, A., Chenthamarakshan, V., Lozano, A., Das, P., & Tang, J. (2023). Protein representation learning by geometric structure pretraining. arXiv. Available at: https://arxiv.org/abs/2203.06125 [Accessed 1 February 2025].

148. Zou, K., Wang, S., Wang, Z., Zou, H., & Yang, F. (2023). Dual-Signal Feature Spaces Map Protein Subcellular Locations Based on Immunohistochemistry Image and Protein Sequence. *Sensors, 23*(17), 9014. https://doi.org/10.3390/s23179014

9 Appendix A

9.1 Ethics approval

UNIVERSITY OF CAPE TOWN
Faculty of Health Sciences
Human Research Ethics Committee

Room 45 E-52-E-Floor- Old Main Building
Groote Schuur Hospital
Observatory 7925
Telephone [021] 406 6492
Email: hrec-submissions@uct.ac.za
Website: www.health.uct.ac.za/home/human-research-ethics

07 August 2023

HREC REF: 557/2023

Dr S Musalula
Department of integrative Biomedical Science
Computational Biology office Site
Email: musalula.sinkala@uct.ac.za
Student: MSPSIB002@myuct.ac.za

Dear Dr Musalula

PROJECT TITLE: COMPARATIVE EVALUATION OF CONVOLUTIONAL NEURAL NETWORK AND TRANSFORMER MODEL FOR PREDICTING PROTEIN LOCALIZATION IN EUKARYOTIC CELLS THROUGH IMAGE ANYALYSIS
(MSc (Med) BIOINFORMATICS - MISS SIBONGISENI MSIPA)

Thank you for submitting your study to the Faculty of Health Sciences Human Research Ethics Committee (HREC) for review.

It is a pleasure to inform you that the HREC has **formally approved** the above-mentioned study.

Approval is granted for one year until the 30 August 2024.

Please submit a progress form, using the standardised Annual Report Form (FHS016) if the study continues beyond the approval period. Please submit a Standard Closure form if the study is completed within the approval period.
(Forms can be found on our website: www.health.uct.ac.za/fhs/research/humanethics/forms)

The HREC acknowledge that the student: Miss Sibongiseni Msipa will also be involved in this study.

Please quote HREC REF 557/2023 in all your correspondence.

Please note that the ongoing ethical conduct of the study remains the responsibility of the principal investigator.

Please note that for all studies approved by the HREC, the principal investigator **must** obtain appropriate institutional approval, where necessary, before the research may occur.

Yours sincerely

Signed by candidate

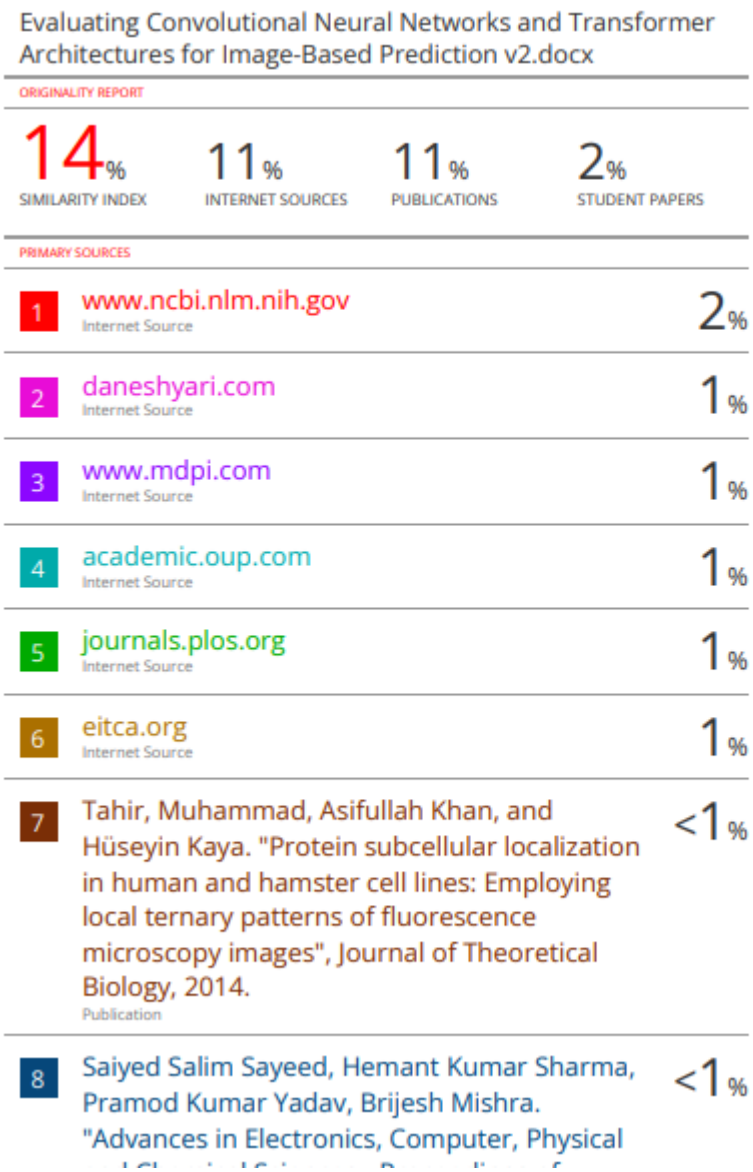
PROFESSOR M BLOCKMAN
CHAIRPERSON, FACULTY OF HEALTH SCIENCES HUMAN RESEARCH ETHICS COMMITTEE

Federal Wide Assurance Number: FWA00001637. Institutional Review Board (IRB) number: IRB00001938 NHREC-registration number: REC-210208-007

HREC/ref: 557.2023

10 Appendix B

10.1 TurnitIn report



Evaluating Convolutional Neural Networks and Transformer Architectures for Image-Based Prediction v2.docx

ORIGINALITY REPORT

14%	11%	11%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.ncbi.nlm.nih.gov Internet Source	2%
2	daneshyari.com Internet Source	1%
3	www.mdpi.com Internet Source	1%
4	academic.oup.com Internet Source	1%
5	journals.plos.org Internet Source	1%
6	eitca.org Internet Source	1%
7	Tahir, Muhammad, Asifullah Khan, and Hüseyin Kaya. "Protein subcellular localization in human and hamster cell lines: Employing local ternary patterns of fluorescence microscopy images", Journal of Theoretical Biology, 2014. Publication	<1%
8	Saiyed Salim Sayeed, Hemant Kumar Sharma, Pramod Kumar Yadav, Brijesh Mishra. "Advances in Electronics, Computer, Physical	<1%