Goldsmiths
UNIVERSITY OF LONDON

Predicting Fossiliferous Localities: a spatio-temporal approach

by

Harris Barra

BSc Computer Science

2025

Declaration

This project report is submitted in partial fulfilment of the requirements for the degree of BSc Computer Science. I declare that this thesis was composed by myself, that the work contained therein is my own, except where explicitly stated otherwise in the text, and that it has not been submitted, in whole or in part, for any other degree or professional qualification.

Harris Barra

Word Count: 9972 words

This thesis was conducted under the supervision of Dr. Tim Blackwell.

Abstract

Palaeontology is a nascent field, where exploration is still deeply rooted in archaic methodologies. Although work has been done in adjacent fields, there is but a growing interest in AI and machine-learning methods to enhance efforts in pinpointing fossil-bearing regions. This study explores the relationships in deep-time environmental data in conjunction with contemporary spatial patterns to produce a model capable of isolating domains of fossiliferous significance by geological eras. Initially by exploring feasibility in proxy data, followed by end-to-end training of neural architectures, furthered by a transfer-learning step and concluding with the merging of spatial and temporal results to produce a final output. The final holistic output produced a Precision of 0.8073, Recall of 0.2606 and AUC-ROC of 0.8569 through cross-validation.

keywords: temporal-model, neural-networks, fossil prediction, deep-time

Dedication

I would like to give thanks to my supervisor, Dr. Tim Blackwell, whose insights and support were invaluable to the study. I also would like to thank my mother for her continuous reassurance and unwavering confidence that inspired me to be the person I am today.

Contents

List of Figures

List of Tables

Abbreviations

ANNs Artificial Neural Networks.

BPPT Back-Propagation Through Time.

CNN Convolutional Neural Network.

CSV Comma-Separated-Value Format.

DEMs Digital Elevation Models.

ENM Ecological Niche Modelling.

GCNs Graph Convolutional Networks.

GIS Geographic Information System.

GRU Gated Recurrent Network.

Ka Kiloannum.

LSTM Long Short-Term Memory.

Ma Megaannum.

Mya Million years ago.

Myr Million year.

NDVI Landsat Normalised Difference Vegetation Index.

NPY Numerical Python Format.

RNNs Recurrent Neural Networks.

Note: Author abbreviations are shown in their corresponding reference entry.

Chapter 1

Introduction

Fossils, the preserved remains or impressions of organisms, are invaluable to the field of palaeontology. Beyond providing taphonomic records of extinct life, they can provide insights into previous environmental conditions through; temperature fluctuations, atmospheric compositions, bathymetric and hypsometric changes. Phenomena that multi-disciplinary research can benefit from including long-term climate trends and future climate predictions (Wang, 2024).

Prospecting typically involves exploiting known fossiliferous outcrops (Anemone et al., 2011), especially sedimentary formations, where chemical interactions between organic and lithological matter enhance preservation (Bogdanovich et al., 2021), as expeditions are high-risk and difficult to fund if returns are uncertain (Hlusko, 2010).

Recently, computational methods like statistical Geographic Information System (GIS) analysis (Oheim, 2007) and machine learning (Malakhov et al., 2009) have been used to evince unseen localities. However, with growing technological advancements, neural models (Anemone et al., 2011) have attempted to absolve serendipity in fossil discovery by mitigating non-linear complexities. Yet, these methods are far from perfect and primarily function as an "exploration filter" (Block et al., 2016), aiding but not fully ameliorating the challenge of reliably identifying fossil hotspots.

As we regress through time, there is increasingly less data and subsequently, fewer predictive studies. Particularly neural-temporal approaches. This gap in deep-time research is overshadowed by simpler methods that do not emphasise or utilise preservation data. This project strives to rectify inhibitions by improving upon limitations imposed by existing work,

validating correctness of existing research and ultimately, by coalescing deep-time preservation patterns, provide a new investigative avenue.

1.1 Objectives and Aims

The fossil record is extensive but heavily concentrated in North America and Europe, with 45% of the world comparatively undersampled (Ye and Peters, 2023). This untapped potential is evidence for prospecting globally which could present further field knowledge.

The primary objective of this study is to quantify the success of a deep-time preservation model on a worldwide basis. By probing macro-level changes we hope to produce results that justify temporal modelling and demonstrate assertive metrics to persuade further exploration.

Due to the ambiguity in the fossil record, we will first establish targets for the study's analysis. These targets, along with a curated series of features, will be fed to a variety of models to analyse non-temporal and temporal significance. We will then elect a tuned architecture for Transfer-learning, allowing us to explore truthful occurrences while preserving geological suitability patterns.

Subsequent performance metrics will inform the development of a supplementary spatial model that focuses on contemporaneous patterns to support potential prospecting. Finally, harmonic performance will be evaluated through superimposed predictions to conclude the study.

Chapter 2

Background Research

Many factors affect fossil likelihood, typically grouped into subcategories for various analyses. These subcategories fall into three domains: Ecological, Preservation, and Discoverability-based. Each offers a unique lens on fossil distribution, highlighting different variables that influence fossiliferous significance. We will explore the advantages and disadvantages of each path in the following literature.

2.1 Machine-Learning and Neural Approaches

The majority of existing work focus on a combination of statistical and machine-learning. The most frequent are discovery models using satellite imagery. A wide array of literature detail fructuous success in surficial exploration models: (Malakhov et al., 2009); (Hlusko, 2010); (d'Oliveira Coelho et al., 2021).

Malakhov et al. (2009) perform spectral analysis using Landsat imagery to identify geologic-specific soils. Although it has merit, their research had a variety of constraints. Their spatial resolution was of 28.5 metres, although very precise, the nature of their task was to understand surficial detail, therefore succumbed to spectral similarity where they were unable to differentiate the sedimentary layer being Cretacean or Paleogenic. They also needed to verify their findings through field verification.

This analysis is improved upon with Anemone et al. (2011), they incorporate Artificial Neural Networks (ANNs) to perform spatial predictions. They refer to the same dataset used by Malakhov et al. (2009), but the latter's usage was for the identification of sedimen-

tary layers instead of exclusively fossil-bearing geologic formations. Anemone et al. (2011) provide a more robust solution involves the ability to classify distinct land coverages, categorisation Malakhov et al. (2009) was not able to achieve. Additionally, they increased spatial resolution by 2× using digital upscaling and they involved a post-process stage where they add classifications rules to improve accuracy. One of the rules they incorporate involves the use of Digital Elevation Models (DEMs), each pixel has its corresponding slope classified as actively eroding depending on whether the slope is $\geqq 5\%$.

Another fundamental constraint is quantifying all fossils in a given area and by extension, true absence. True absence is marginally absolved through preservation and discovery techniques by looking at geological implausibilities, however it can be further mitigated by ecological analysis. Myers et al. (2015) investigates Ecological Niche Modelling (ENM), in which multivariate functions are used to find relationships between occurrences and environmental factors. Presence-only algorithms such as GARP, MaxEnt and PaleoENM are used to obtain statistical probabilities to understand a geographical distributions. However, a prerequisite is having a validated ethological study of the target species, this limits the application to localised examples and in the broader context of life on earth, presents an unfeasible challenge as 99% of evolved life is extinct (Novacek et al., 2001). Figure 2.1 (Hendricks, 2024) illustrates the time-scale of earth.

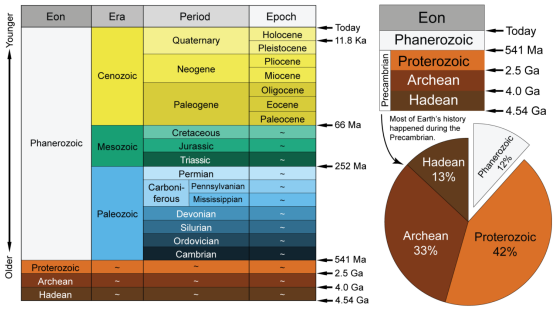

Figure 2.1: Geologic Time Scale (Hendricks, 2024).

Despite this limitation, it can be adapted to work at a class level set of specimens to gather a broader range of reliable information. They bring to light the usage of stratigraphy as a surficial exposure factor. They detail stratigraphic layer recreation, of a prehistoric seaway, using coarse information sourced from other literature.

Phillips et al. (2006) concur with the effectiveness of ecological research through mod-

elling geographic distributions and by comparing the efficiencies of MaxEnt and GARP. Usage of environmental variables show that vegetation indices can improve discovery accuracy in some cases. One caveat to note, is that genera from the study persist in the current-day, thus we can fully verify true absence by the conditions of their environments. This cannot be reliably extrapolated for deeper-time genera as it becomes decreasingly serendipitous.

2.2 GIS and Statistical Approaches

Oheim (2007) investigates discovery potential using contemporary data. They investigate correlations between 4 variables: geology, elevation, vegetation and distance-to-roads. Using a limited number of variables and refinement through field-work they achieved a "statistically significant 90% density variation". Performance is evidence for the intrinsic simplicity of discovery models and therefore the avoidance in solving the complexities of taphonomic-focused models. This results in inability to transfer the model to forecast on other regions, contextual knowledge is explicit to this one specific formation and there is no further investigations whether more geo-spatial variables are required depending on the characteristics of other formations.

A core inspiration for our study stems from Block et al. (2016). Their study entails superimposing 3 statistical models, each of which used mixture of general linear and logistic models to address different predictive categories as illustrated in Figure 2.2 (Block et al., 2016). One of which propagates shallow-time (0-120 Kiloannum (Ka)) history to build taxon-specific ecological suitability. This is particularly intriguing as there is no mention or suggestion for a neural approach in either preservation or ecological models, arguably due to the over-reliance of subject genera ecology, which we explored with Myers et al. (2015). However, Block et al. (2016) mention how their predictions could be improved through more detailed data and increased spatial resolution. We can take a new angle on this by removing genera from the equation and analyse deep-time preservation to the fossil record as a whole.

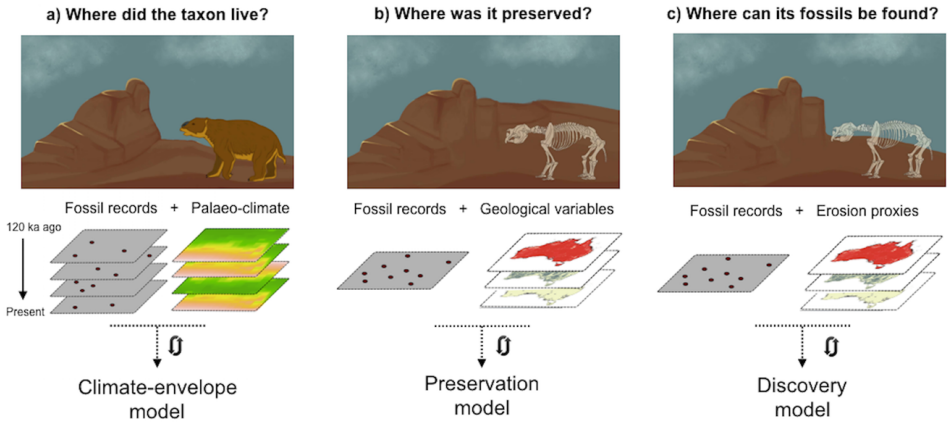

Figure 2.2: Combined Models Approach (Block et al., 2016).

2.3 Limitations in Existing Work

A consistent limitation is the fidelity of data that can be found, often the data has insufficient resolution or proxies derived from sparse information (see Figure 2.3 (Block et al., 2016)).

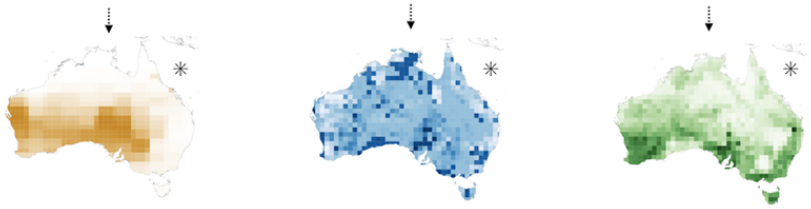

Figure 2.3: Sequence of $1 \times 1°$ Models (Block et al., 2016).

Data in every model approach is limited. While tighter spatial focus offers higher data availability, making them ideal for case studies, scaling beyond these areas presents challenges. For instance, Wills et al. (2017) discusses the potential for multi-continental expansion but does not explicitly address the feasibility of acquiring the necessary prerequisite data or the suitability of models for forecasting these unseen outcrops.

Discovery models are trained to work effectively on the environments of test cases and therefore models trained on an arid landscape in Australia will not accurately predict on a rainforest in South America where data reconnaissance poses a larger challenge. Although relation does not infer causation, there is room to study at a global frontier without imposing training to regional confinements.

Contemporaneous data presents fewer limitations compared to historical records. This is

evident in studies like Block et al. (2016), which focus on the relatively recent Quaternary period within the eon of the Phanerozoic (see Figure 2.1). The recency of this period allows models to leverage a richer dataset, whereas older geologic periods face more significant data constraints.

However, there does exist predictive data for the Mesozoic (Scotese et al., 2024*b*) using the same HadCM3 model which could be used in a similar fashion to the Pleistocene data in Block et al. (2016) to fuel rich taphonomic, though limited ecological inferences.

Temporal constraints also pose challenges such as subjecting genera to a larger stratigraphic/periodic interval. Relevant information can be lost at the chance of larger datasets. The same can be said conversely where a smaller interval will retain more information, but reduce the dataset. Naturally, this is causation for extrapolation. Block et al. (2016) illustrates some of the best temporal confinements, a short temporal focus and a dataset of 10,000 samples, but still succumbs to issues such as low spatial resolution.

Wills et al. (2017) and Block et al. (2016) have differing conclusions about the combined use of presence data and background data. The former suggests that due to the overarching issues of not knowing true absence, adequate background data can be used in conjunction with presence data to eliminate bias. On the contrary, the latter interpret the output as a non-binary ranking of suitability. Given the sensitivity around using specific environment variables, knowledge of genera, and the impossibility of knowing how many fossils can be found, Block et al. (2016) suggests a more responsible choice by not enforcing binary constraints from the model and only after when applying an evaluation to maximise potential in the favour of the study.

In regards to the aims of the project, we've explored the impracticalities of ecology in the broad view of extinct life. Fossil analysis itself is constrained by fossil availability, creating a paradox where the lack of fossil data causes ecology to remain speculative. Yet without robust modelling, fossil discovery and interpretation remain limited. This cyclical dependency accentuates the balance between scientific inference and reliable prospecting.

The constructive approach is to explore broad preservation and discovery to disassociate taxonomic variability but not prevent fossil classification if future studies wish for further granularity.

2.4 Temporal Neural Networks: RNNs

While ANNs have proven effective for discovery models (Anemone et al., 2011), they face two significant challenges in historical modelling: temporality and vanishing gradients.

Recurrent Neural Networks (RNNs) have become a fundamental architecture in deep-learning for handling sequential data. Unlike traditional feed-forward networks such as ANNs, RNNs are specifically designed to capture temporal dependencies and address these issues.

2.4.1 Theoretics of RNNs

RNNs differ architecturally from feed-forward networks by establishing a feedback loop (Stryker, 2024) to contextualise information across multiple sequences (see Figure 2.4).

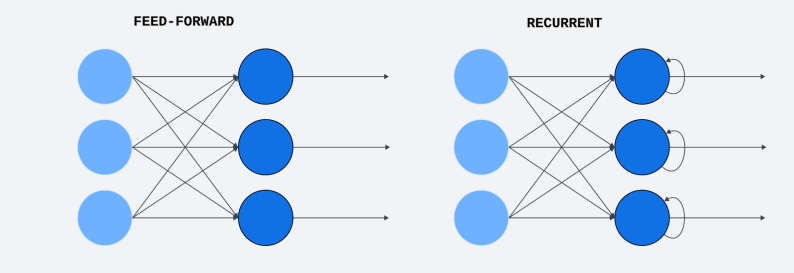

Figure 2.4: Feed-forward vs. Recurrent Network.

Traditional models handle inputs independently, missing mechanisms to account for prior inputs. RNNs employ Back-Propagation Through Time (BPPT) for gradient calculation, akin to standard backpropagation but modified for sequences. BPPT also aggregates errors over time via shared parameters (Stryker, 2024), thus allowing temporal learning.

However, even simple RNNs face the vanishing gradient issue, where continuous BPPT gradient multiplication exponentially minimises weights, hindering long-term dependency retention.

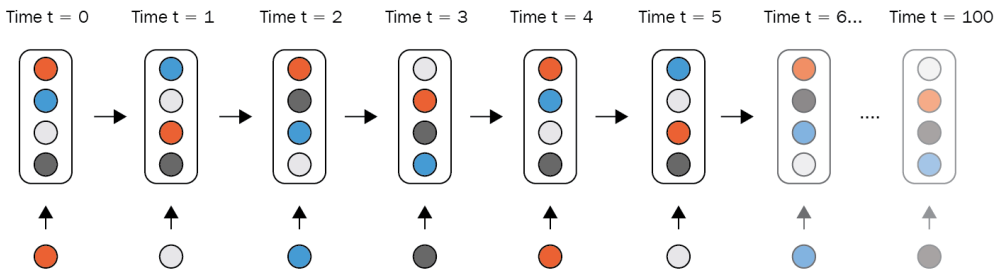

Figure 2.5: Example of Vanishing Gradient (Karim and Menshawy, 2018).

Naïve solutions worsen the vanishing gradient problem through increased neurons or layers. Advanced RNNs (Gated Recurrent Network (GRU), Long Short-Term Memory (LSTM)) use gates to selectively retain and update long-term contextual information.

2.4.2 Long-Short Term Memory (LSTM)

LSTMs were introduced by Hochreiter and Schmidhuber (1997) to address vanishing gradients and tackle this by incorporating a memory cell, a hidden state and three gates: *input*, *forget*, and *output*.

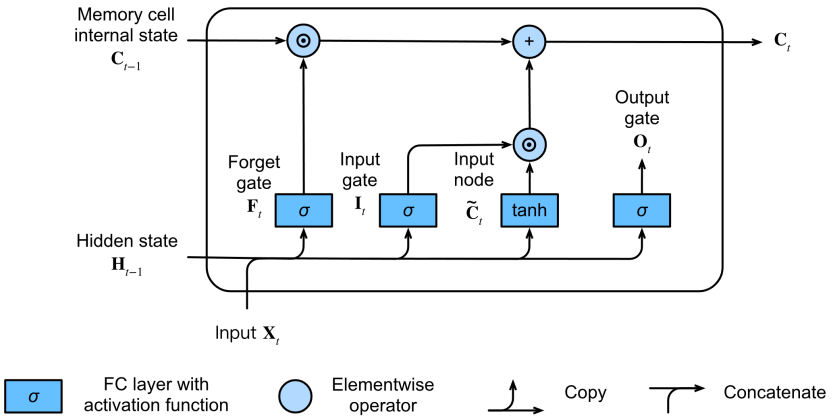

Figure 2.6: LSTM Unit with Internal State (Zhang et al., 2021).

The input node ($\tilde{C}_t$) represents a candidate update for the memory cell and is computed similarly to the gates but uses a Tanh (Hyperbolic Tangent) function to constrain the interval between (-1, 1).

$$\tilde{C}_t = \tanh(W_C \cdot [H_{t-1}, X_t] + b_C) \tag{2.1}$$

The memory cell state is updated by combining past memory and the new candidate memory, weighted by the *forget* and *input* gates:

$$C_t = F_t \odot C_{t-1} + I_t \odot \tilde{C}_t \tag{2.2}$$

This is the mechanism that helps mitigate the vanishing gradient problem (Zhang et al., 2021).

2.4.3 Gated Recurrent Network (GRU)

GRUs follow a similar architectural pattern; a hidden state, instead of three gates, there are two. They also contain a candidate hidden state. Each of these are computed distinctly and work in tandem to control the flow of information.

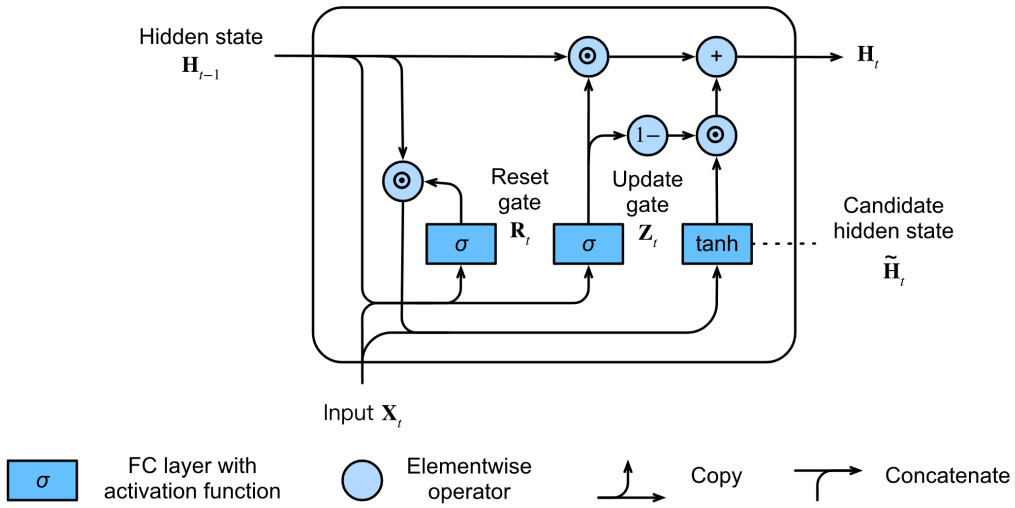

Figure 2.7: GRU Unit (Zhang et al., 2021).

Figure 2.7 (Zhang et al., 2021) illustrates how the complete unit processes information.

- $\tilde{H}_t$ (Candidate Hidden) is computed using the *reset* gate (R_t).

$$\tilde{H}_t = \tanh(W_{\tilde{H}} \cdot [R_t \odot H_{t-1}, X_t] + b_{\tilde{H}}) \tag{2.3}$$

- H_t (Hidden) is computed by combining the previous hidden state and the candidate hidden state, using the *update* gate (Z_t).

$$H_t = Z_t \odot H_{t-1} + (1 - Z_t) \odot \tilde{H}_t \tag{2.4}$$

Unlike LSTMs, which use separate memory cells and hidden states, GRUs merge these roles into a single hidden state, making them more computationally efficient. This design allows GRUs to adapt and capture both short-term and long-term dependencies (Zhang et al., 2021) in sequential data while reducing computational overhead compared to LSTMs.

Chapter 3

Methods

This section details all the methodological choices that were made and evolved throughout the course of the project. The first section focuses on architectures, followed by data reconnaissance and the last section details evaluation methodology.

3.1 System Design

As illustrated in Figure 3.1, the proposed system can be divided into three core sections: Pre-Processing, Models, Post-Processing/Evaluation.

Pre-Processing (Section 3.3) will primarily involve incorporating a variety of features, performing consistency remedies, producing appropriate targets.

Model approaches will be discussed (Section 3.2) for both Preservation and Discovery models. We will explore the inherent benefits in the model choices and their potential limitations with complementing solutions.

Lastly we will perform three sets of evaluations (Section 3.5), each of which are to investigate validity, reliability and suitability in the pipeline. We will evaluate the temporal model on its own to build an understanding of the temporal-neural behaviour, evaluation of the spatial model will help us understand impact and further neural behaviour, the subsequent harmonic model (Section 3.4) evaluation will conclude our analysis.

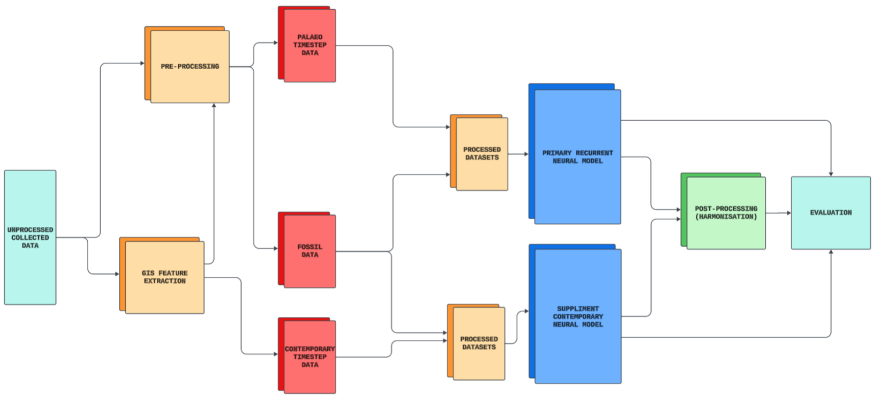

Figure 3.1: Proposed System.

3.2 Model Design

Before we can explore targets to fit to the models, we firstly need to construct an understanding of the models to ensure appropriate utilisation.

3.2.1 RNN Architecture Choices

We explored the standard temporal architectures for deep-learning in our literature review (see Section 2.4). The primary model will leverage LSTMs as they are designed to capture temporal dependencies and patterns the best, however we will also investigate significance between GRUs and LSTMs to determine long and short term information reliability. In our case we will study environmental and geological variables over time for each unit of land. This design ensures that even subtle temporal shifts in variables like precipitation or temperature are preserved and contribute meaningfully to fossil suitability predictions.

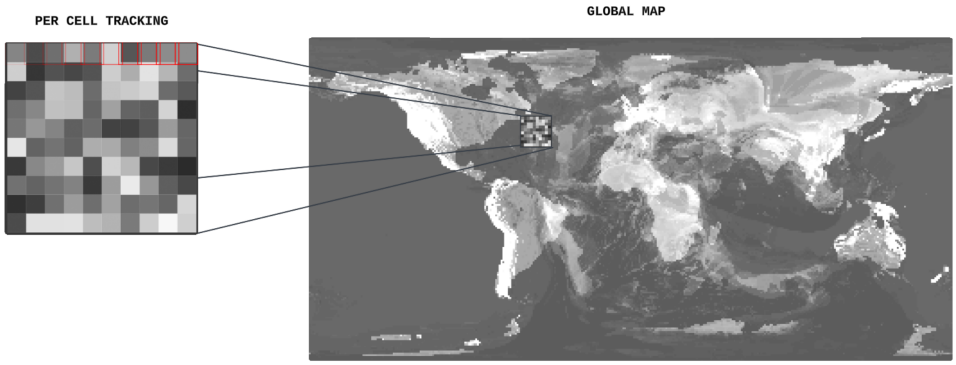

Figure 3.2: Visualisation of Cellular Map.

Fossil preservation is often influenced by highly localised factors, such as intense erosion or sedimentary deposition (Block et al., 2016), which vary significantly over time. Analysing each cell independently (see Figure 3.2) across time ensures that these granular variations are captured, preserving the fidelity of temporal trends that may otherwise be lost in spatial aggregation.

3.2.2 RNN Staged Strategy

An initial prototype to the preservation model was to feed a geologic specific (Triassic) set of sedimentary occurrences and observe if a simple temporal model gathers correct patterns. Unfortunately, the complexities of taphonomy prevented any successful measure from the prediction as shown by the output in Figure 3.3.

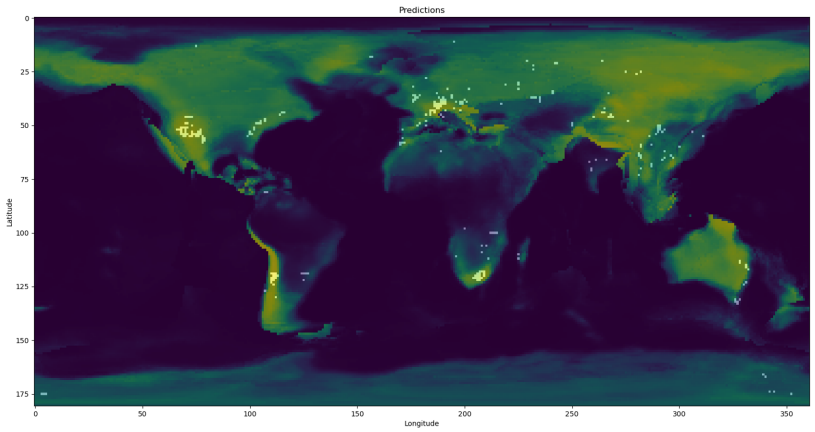

Figure 3.3: $1 \times 1°$ Prototype Predictions demonstrating severe misclassification (Yellow = High Confidence, Purple = No Confidence).

Upon close inspection we can determine an understanding of the behaviour. The prediction is ambiguous and rarely illustrates bias towards known cluster locations, it makes wide predictions across different continents reflecting unrealistic spatial accuracy, additionally predicting incredibly implausible locations such as metamorphic, volcanic and igneous bedrock, with high confidence. Therefore, a more fundamental approach needs to be taken to ensure that the model can understand core principles.

For our experimental pipeline, we will begin by understanding the fundamentals of the elected approximation data (see subsection 3.3.2) to determine viability, we will then approach predictions on a categorical level before finally Transfer-learning against known true occurrences in the final stage. Figure 3.4 illustrates this.

1. *Stage 2* (4.2): Analysis on a suite of features to determine suitability of temporality by comparing various models.

2. *Stage 3* (4.3): Predicting geologic classes against a target whereby true absence is completely absolved.

3. *Stage 4* (4.4): Transfer-learning on the *Stage 3* model to reincorporate absence.

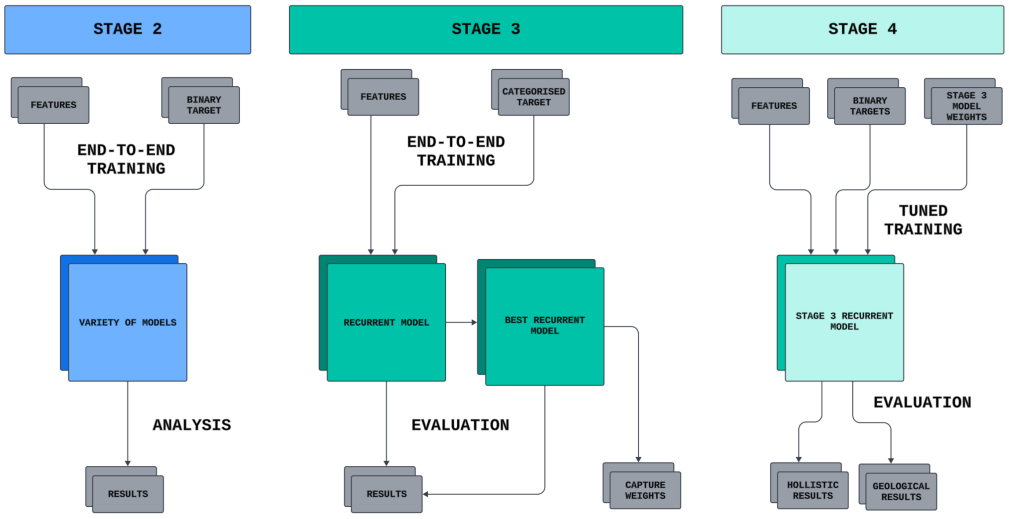

Figure 3.4: Overview of the three-staged pipeline.

3.2.3 Spatial Architecture Choices

Our spatial model's purpose is to produce localised predictions by aggregating neighbouring cellular information. Anemone et al. (2011) attempt their model through an ANN due to

its robustness in solving simple patterns. We will however attempt to explore these relations through a Convolutional Neural Network (CNN) as they are often considered better for spatial tasks due to their architectural design (Idrees, 2024). Convolutional layers are designed to work with structured, grid-like images (Idrees, 2024). They do this by using convolutional filters that slide over the input vectors, detecting patterns and features at different locations, see Figure 3.5. This is particularly relevant for identifying fossiliferous localities, as a cluster of independent signatures could be crucial for spatial inferences.

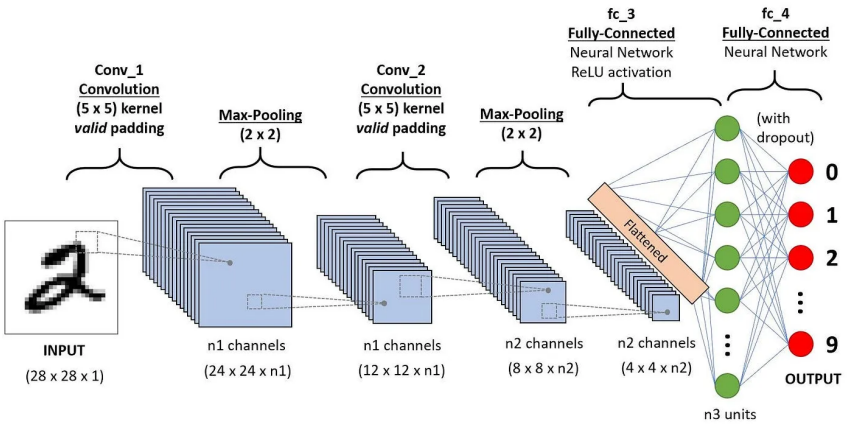

Figure 3.5: Overview of CNN Image Recognition (Joseph and Elleithy, 2020).

ANNs typically treat each feature (e.g., the spectral bands of a pixel (Anemone et al., 2011)) independently without explicitly considering the spatial relationships between neighbouring pixels. While an ANN can learn complex relationships, it requires features to be explicitly engineered or learning implicitly through a large number of neural connections and training samples, which can be less efficient and in our approach, counter-intuitive.

There are other spatially-conforming architectures such as Graph Convolutional Networks (GCNs) to potentially enlist, however to maximise compatibility and efficiency, we will be explicitly implementing a CNN model to utilise a shared list of features in the study.

3.3 Dataset and Prediction Targets (Pre-Processing)

Now that we've concluded the appropriate models for the investigation, we need to compile features and deriving our own targets. Since taphonomy is shaped by multiple processes, a pre-defined set of ambiguous features may not capture the nuances necessary for prediction.

We must curate our own selection of features that specifically emphasise taphonomic controls. This includes variables related to burial, post-burial diagenesis and weathering - all of which contribute to the selective preservation of fossils.

Recall that purely predicting fossils cannot usually be defined true or false labels due to uncertainty in true/false absence. This means framing our targets in a way that captures relative importance rather than striving to achieve the impossibility of absolute correctness, in doing so, we aim to enhance our model's ability to detect meaningful correlations within deep-time data, thus improving our understanding of how and where fossils are preserved.

3.3.1 Establishing Ground Truth Targets

Sedimentary vs. Non-Sedimentary

We start by identifying the regions of interest, fundamentally sedimentary against non-sedimentary regions. This is an exploration step to investigate the results on a grand scale without enforcing granular recognition learning (see Section 4.2). There are limited but few options for global-spanning lithological maps.

The original reconnaissance point was Macrostrat, as it is the largest platform of aggregated and distributed geological map data.

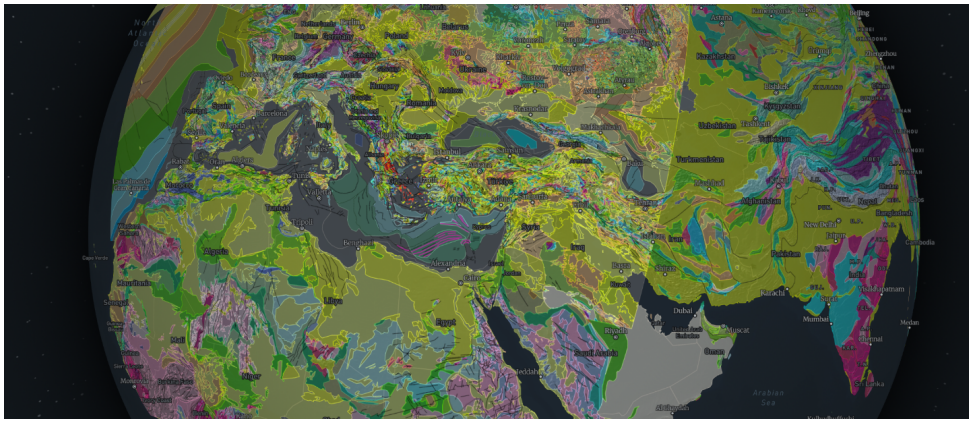

Figure 3.6: Snapshot of Bedrock Surveys from Macrostrat (2025).

Due to the aggregative nature of the site, there are considerable resolution inconsistencies between different sources (highlighted in Figure 3.6). One could opt to produce a map at a higher-resolution but substantial work would be needed to keep fidelity consistent.

Through metadata inspection we were able to isolate a consistent single resource, Chorl-

ton (2007), which provides even global distribution. The source states that this data is compiled from various sources and disclaims expected spatial inaccuracy, however this is not a serious issue for predictions at our working-space resolution and consistent geographic distribution is necessary for fair predictions and learning evaluations.

This data is imported into QGIS (GIS Software) where regions are extracted based on the survey's designated lithology of each polygon. We can then group these lithologies to produce a map of sedimentary polygons as shown in Figure 3.7.

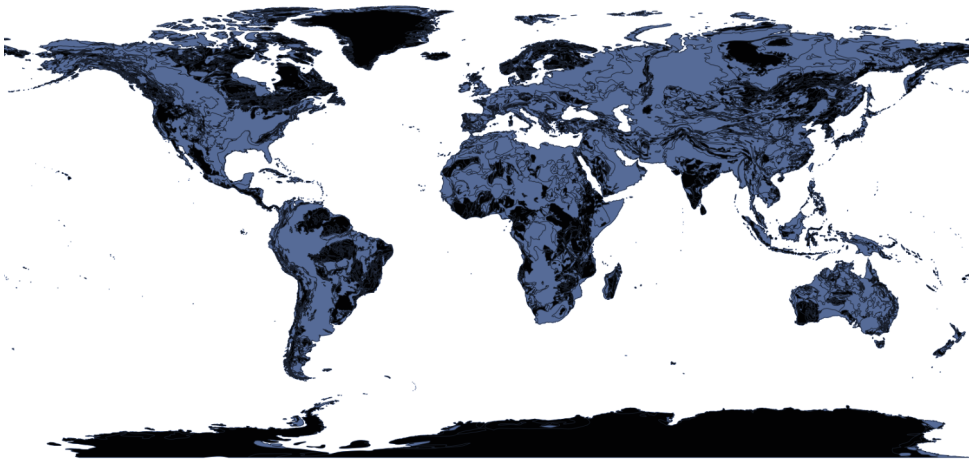

Figure 3.7: Sedimentary Map. Blue = Sedimentary, Black = Non-Sedimentary (e.g., volcanic, Igneous , etc.).

Cleaning The Fossil Record

Before we can build any further targets, specifically ones inferred through fossils, we must scrub occurrence data as the fossil record is convoluted and unsanitary. We dispatched queries to Paleobiology Database to obtain fossils within a 541 Myr-scope. They offer an array of filters for researchers to customise reconnaissance (see Table A.6).

	Min / Max Range (Ma)	**Non-Marine Occurrence Count**	**Marine Occurrence Count**
Paleogene	23 – 66	48583	107195
Cretaceous	66 - 145	46974	156020
Jurassic	145 - 201	12631	133246
Triassic	201 - 252	8816	74660
Permian	252 - 299	8876	113327
Carboniferous	299 - 359	12208	60959
Devonian	359 - 419	3864	84993
Pre-Devonian	419 - 549	2542	213839

Table 3.1: Occurrence counts.

Table 3.1 presents the occurrence numbers before we perform any sanitisation. Data was cleaned by eliminating entries which did not have valid: min-max, coordinates or environment bins.

Periods are not simply given discrete age intervals, there is a natural progression into newer periods, thus we define our hard-limit by allowing a ± 2 Myr buffer to accommodate for sediment movement and any record inaccuracies.

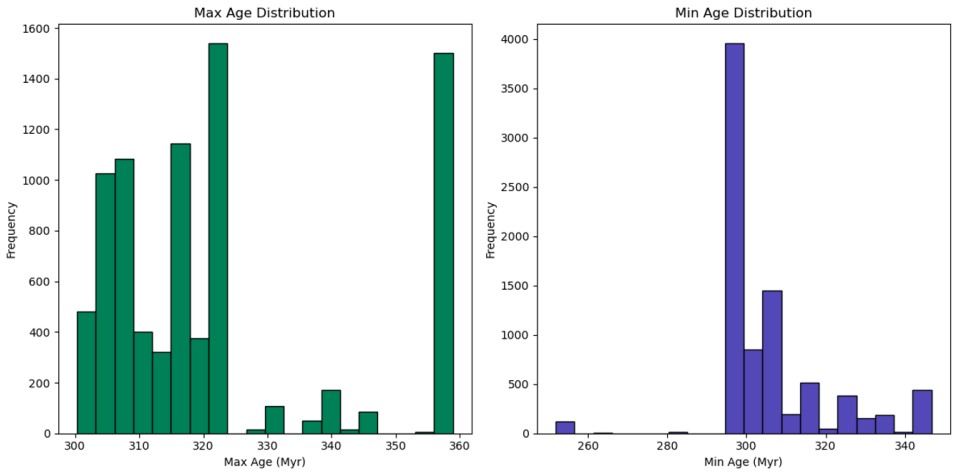

Figure 3.8: Carboniferous occurrences before cleaning.

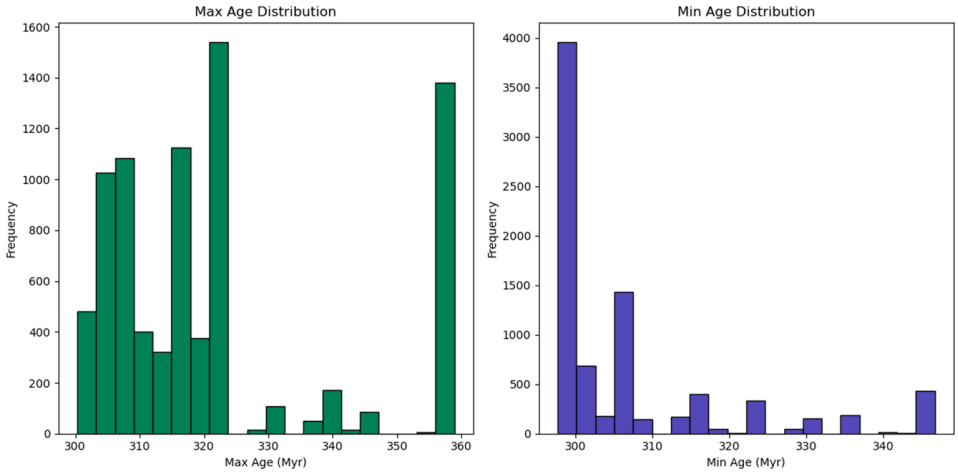

Figure 3.9: Carboniferous occurrences after a ±2 Myr buffer.

	Non-Marine Percentage Removed	**Marine Percentage Removed**
Paleogene	8.19%	2.48%
Cretaceous	13.70%	2.02%
Jurassic	16.02%	1.58%
Triassic	8.94%	1.12%
Permian	12.21%	0.98%
Carboniferous	32.95%	0.80%
Devonian	41.15%	2.29%
Pre-Devonian	93.15%	1.37%

Table 3.2: Removal percentages.

Non-Marine tend to have much more indeterminate ages, likely due to high energy conditions. Another trend is that older fossils are typically more imputable, however this is mostly due to the environment type not being recorded.

We also focus our fossils from 23 Ma onwards, following the findings of Ye and Peters (2023) which highlights how fossils from the late-to-current Cenozoic can complicate sediment dating. In the next subsection, we aim to perform our own dating process and therefore pre-emptively address this issue by excluding fossils that date between 0-23 Ma.

Aged Stratigraphic Polygons

Our next target will now fuse our fossil and lithological datasets. We want to produce a target which can focus on preservation without being constrained to true absence. The solution we propose is to approximate surficial age, where each area is classified on the preservation

likelihood of a specific time-period. In other words, if an area has been subjected to predictable conditions (e.g., weathering), it should reveal the stratigraphic layer that contain fossils matching its age. This allows us to evaluate the preservation potential relative to their expected age.

Once again, we utilise QGIS to overlay occurrences onto sedimentary polygons. Each polygon is then assigned a mean minimum-maximum age range based on the fossils it contains, if at all. Finally, we calculate the 75th percentile average to estimate an approximate regional age.

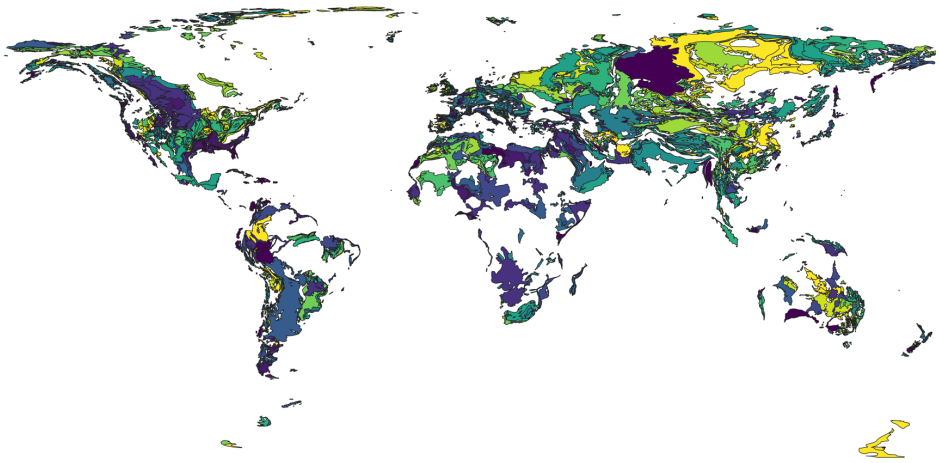

Figure 3.10: Observed Aged Regions Classified by Occurrence Density Ages (Purple = Youngest, Yellow = Oldest).

3.3.2 Approximation Data

We incorporated two main resources, the first resource is the supplementary material (Scotese et al., 2024*a*) from Scotese et al. (2024*b*). This research details the production of palaeo-approximation data through various methods and simulations, the second is through an earlier version of the palaeogeographic data (Scotese and Wright, 2018) but explored further with physiographic derivations (Salles et al., 2022). We wished to author our own extractions through GIS software, but to expedite implementation, and to not repeat existing work, we utilised the following features from each resource:

Features from Scotese et al. (2024a)	**Features from Salles et al. (2022)**
Precipitation	Water Flux
Elevation	Sediment Flux
Temperature	Flood Basins (Lakes)
-	Erosion Rate
-	Uplift Rate
-	Slope

Table 3.3: Extracted features.

These features we intentionally chosen as they influence climatic and environmental macro-changes which caters to our aim of preservative momentousness (see Table A.7). Each feature was extracted at the nearest 5th Ma in the inclusive range of (0-245 Ma), thus for our analysis we will be exploring 50 time-steps of historic data. None of the features were at matching resolutions nor appropriately cleaned, we will discuss consistency and imputation in the coming subsection 3.3.3.

In regards to the spatial dataset, it will utilise the most recent time-step (0 Ma) from the aforementioned resources, however we will also take the opportunity to include Landsat Normalised Difference Vegetation Index (NDVI) to improve surficial exposure as demonstrated in literature (Anemone et al., 2011), (Malakhov et al., 2009).

Extra Feature Extractions

As mentioned in the previous subsection, we wished to obtain our own extractions however considerable analytical effort had already been applied for the existing data, therefore any extra extractions we could obtain ourselves. Figure 3.11 shows a thumbnail of the working environment in QIGS (GIS Software).

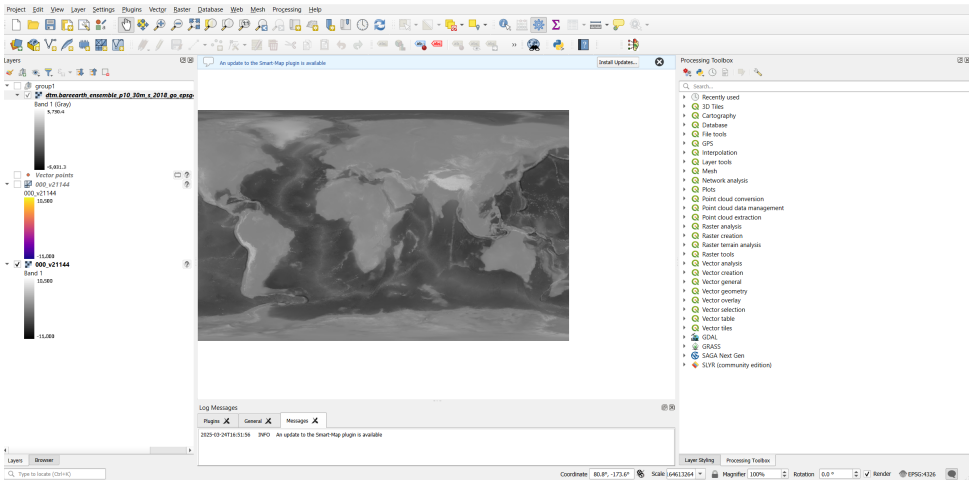

Figure 3.11: Snapshot of QGIS.

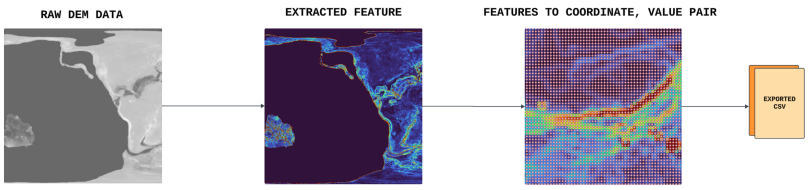

Figure 3.12: Extraction Pipeline.

Figure 3.12 visualises the pipeline we employ to do extractions. From the catalogue of existing features, we warranted just one additional DEM extraction, Flow Accumulation, to further enumerate hydrological features.

One last feature we want to investigate is 'Death Signal', this feature takes our cleaned fossil data (see subsection 3.3.1) and at each time-step we highlight the location of where the supposed occurrence had fossilised in its early diagenesis stage. We produced this through a Python tool which utilises palaeographic transformations (see subsection 3.3.4) to project fossils to the time of their respective deaths. Naturally the rasters are spatially aggregated to the working-resolution, hence they have consistent resolution and are not further imputed/re-scaled.

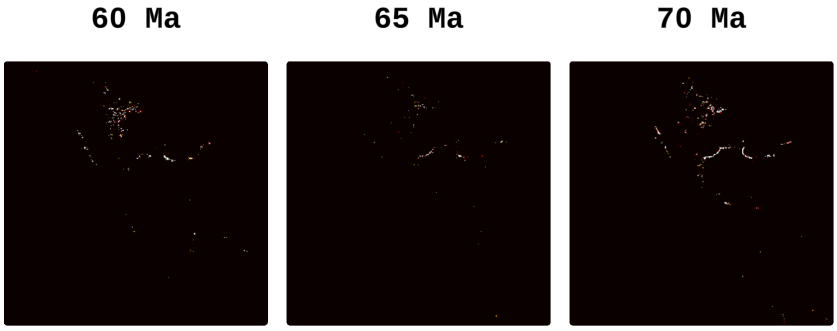

Figure 3.13: Visualisation of Death Signals between 60Ma-70Ma.

3.3.3 Resolution Consistency Strategy

Resampling Techniques

Several techniques were employed to normalise the complete dataset resolution to 0.25 × 0.25°, we used a variety of sampling algorithms for different features to preserve feature fidelity and prevent spatial noise. Despite the emphasis for cellular independence in RNNs, attempting to use a spatially destructive algorithm can indirectly disrupt and confuse training and evaluation as we damage the integrity of the feature as indicated by Figure 3.14.

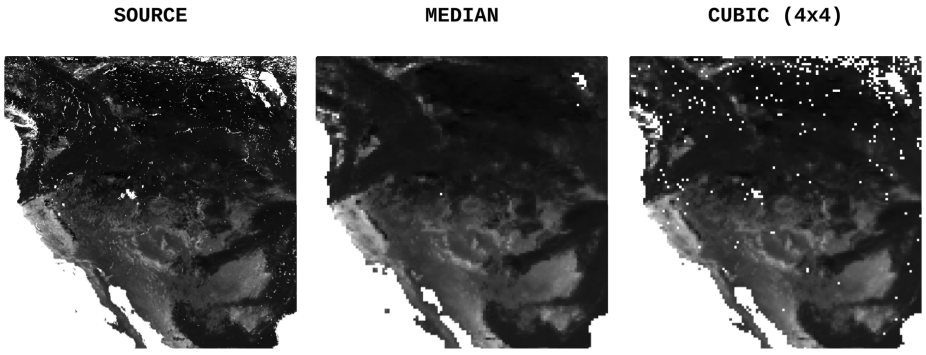

Figure 3.14: Visualisation of NDVI downsampling: candidate preserving algorithm (Median) vs. noise-inducing algorithm (Cubic 4x4).

Feature	Scaling	Sampling Algorithm
Precipitation	Upsample	Kriging
Elevation	Downsample	Median
Temperature	Upsample	Kriging
Water Flux	Downsample	Cubic (4x4)
Sediment Flux	Downsample	Lanczos (6x6)
Flood Basins (Lakes)	Downsample	Median
Erosion Rate	Downsample	Median
Uplift Rate	Downsample	Median
Slope	Downsample	Mode
Flow Accumulation	Downsample	Median
Death Signal (Death-Densities)	N/A	N/A
NDVI	Downsample	Median

Table 3.4: Features against scaling and proposed best sampling algorithm.

Imputation Techniques

Imputation was typically simple, we did not need to indulge in expensive algorithms such as Kriging to fill missing cells although this is something that could potentially be explored for select features (e.g., NDVI). The most frequent method was to grid fill oceanic territories with zero values (see Figure 3.15).

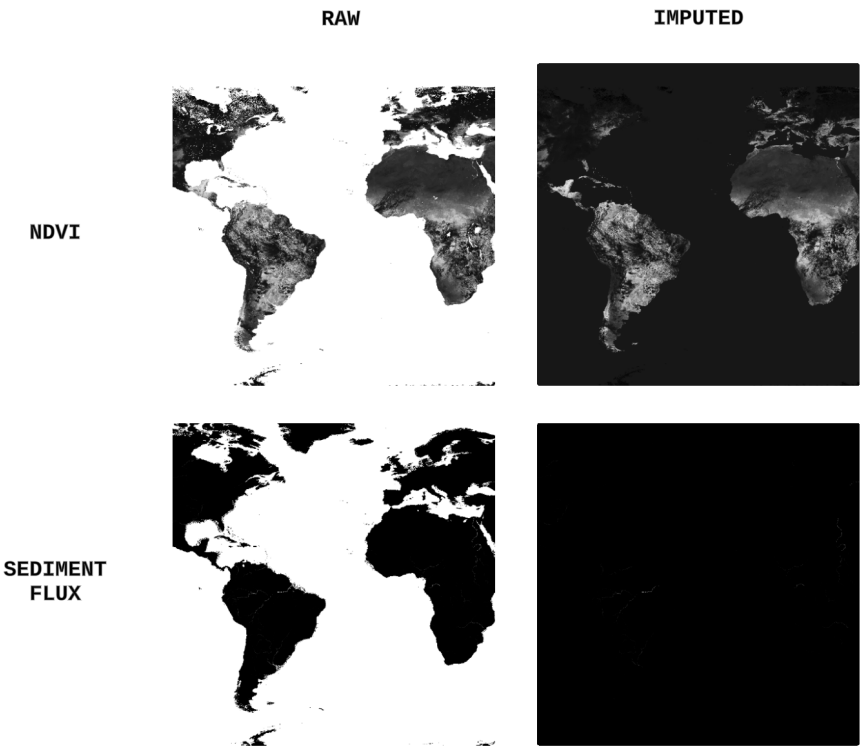

Figure 3.15: Imputation of NDVI and Sediment Flux through zero-ing oceanic regions.

This method is exclusively to retain consistency and not to introduce proposed oceanic values, as we will later note in the implementation chapter, we will not be processing any oceanic/water-body regions to allow us to conduct our research on landmass only samples.

Once features had correct temporal and spatial alignment, we exported them in the same fashion as the latter part of the extraction pipeline (see Figure 3.12). We then converted the exported Comma-Separated-Value Format (CSV)s. into Numerical Python Format (NPY)s using our automation tool (A.1) in Python.

3.3.4 Palaeographic Transformations

RNNs are designed to handle temporal data with no concern for spatial awareness. Due to the large time-scope, RNNs inherently cannot adapt movement a given cell may experience. This is prevalent in our study as the Earth has evolved through many land distributive processes.

The solution is to track each cell at each time-interval to ensure we are building a deep time history of fixed locations. This is achieved using a rotation model tool. We can provide a Python library (GPlates) with coordinates, model variant and time-interval to produce the corresponding palaeo-coordinates. Various rotation models exist, including (Seton et al., 2012), (Wright et al., 2013), and (Torsvik et al., 2012). However, we employ the model by Matthews et al. (2016) due to its rigorous validation on our palaeo-data (Scotese et al., 2024*a*).

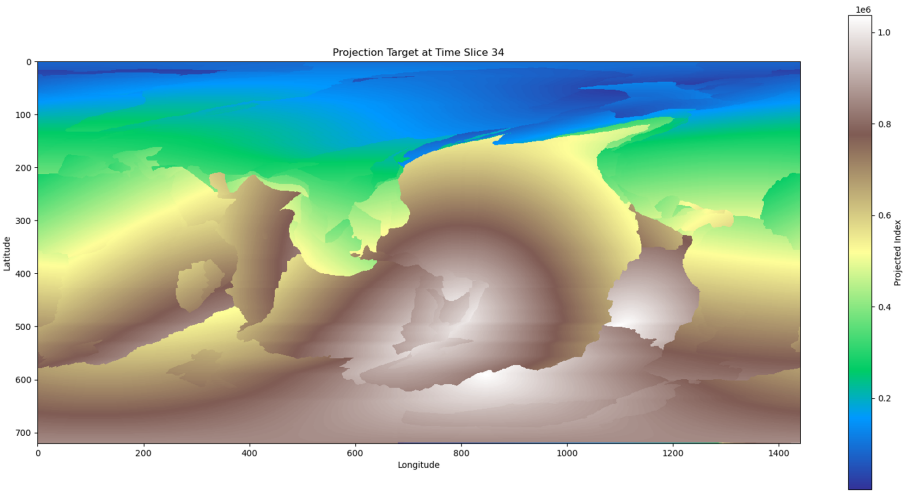

Figure 3.16: Example Projection Matrix at 170 Ma.

Depending on resolution, this can become extremely computationally exhaustive; simply

transforming a 50 time-step dataset at $0.25 \times 0.25°$ would take a theoretical 30 hours. To save on computation time, we can use our Python tool to pre-calculate projections into matrices for each time step. Each index's value in the matrix will point to the corresponding palaeo-coordinate index, this way we can generate a set of indices for each time step which all features can simultaneously attain. Although at $0.25 \times 0.25°$ it takes $\approx$ 3 hours to generate matrices, pre-loading at runtime reduces the compilation to $\approx$ 3 minutes.

3.4 Harmonisation (Post-Processing)

Our final step is the unification of the temporal and spatial models. Our initial designs presented mechanics involving individual weighting and stratigraphy scoring (see A.10), however upon revising the experimental approach we needed to modify this step to better fit into the study.

Instead of a weighting mechanic whereby both the temporal and spatial model are weighted individually, we will superimpose the spatial predictions with a scalar weight element-wise on the temporal predictions (Figure 3.17), this ensures that the temporal model is the dominant model and the spatial behaves as a supplementary and not competitive such as the pipeline in (Block et al., 2016).

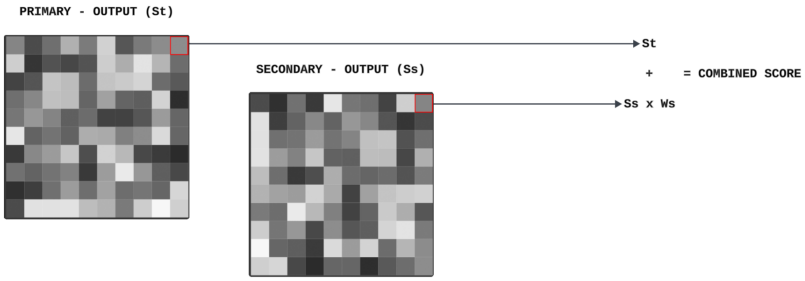

Figure 3.17: Revised weighting mechanic (St = Suitability Temporal, Ss = Suitability Spatial, Ws = Weight Spatial).

3.5 Evaluation Design

We will evaluate our models through a variety of metrics to understand behaviour and analytical inferences, we will also explore two validation methodologies, model tuning and our research questions.

3.5.1 Standard Performance Metrics

Behaviour can be investigated through standard classification metrics such as Accuracy, Precision, Recall, F1-Score, AUC-ROC.

- (Classification) *Accuracy*

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

- *Precision*

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.2}$$

- *Recall*

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.3}$$

- *F1-Score*

$$\text{F1-Score} = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.4}$$

Metric	**Description**
Accuracy	Classification accuracy explains macro-level correctness of class predictions.
Precision	Ensures that predicted areas are reliable by analysing the false positivity of predictions.
Recall	Captures the proportion of correct predictions, ensuring that known and potential fossil sites are not overlooked.
F1-Score	Harmonic mean of Precision and Recall.
Area Under the Curve (AUC-ROC)	Measures diagnostic performance by evaluating the model's fit to random chance.

Table 3.5: Metrics and their behavioural insights.

3.5.2 Confusion Matrix

A Confusion Matrix is a tabular visualisation of classification performance. It shows numeric results to help better understand where a model is erring. Metrics such as Precision/Recall can be derived from confusion matrices however they are less useful in multi-class models

where empirical representations may not be enough to understand where the model may make mistakes.

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Table 3.6: Confusion Matrix.

3.5.3 Threshold-Moving

Threshold-moving is a method used to change the decision-threshold of metrics to address class imbalances. In a study such as this, there will inherently be an imbalance between different categories or even in a binary sense (e.g., sediment vs. non-sediment regions).

Although we could absolve this through over-sampling the smaller class (e.g., SMOTE), we would be doing an injustice to model learning as we reduce total samples. Threshold-moving will be used on our standard metrics to allow us to adjust the required confidence in the model to produce bias primarily in Precision/Recall.

3.5.4 (Fossiliferous) Confidence Metric

Threshold-moving will give us stronger metrics however we also want a metric to present a fundamental confidence in the model. It's imperative to find trustworthy confidence levels to show evidence for a correct prediction with unknown ground truths. This is calculated as follows:

- P_c is the predicted confidence of a given cell c.

- G_c is the ground truth of a given cell c.

$$\text{Confidence Metric} = \frac{\sum(P_c \cdot G_c)}{\sum P_c} \tag{3.5}$$

3.5.5 Spatial Accuracy Metric

Inspired by methods in Oheim (2007), Wills et al. (2017), we wish to quantify spatial predictability. Moran's I investigates spatial autocorrelation but it does not infer spatial extent. Although clustering is importance, our study does not require it as we are limited by spatial resolution and are not investigating ecology.

Autocorrelation describes if the system produces realistic clustering patterns whereas spatial extent demonstrates whether the predicted areas align with known fossil locations. Thus we will measure the overlap between the predicted area and the ground truth area, also called Intersection over Union, or IoU. This is calculated as follows:

- P_c is the predicted confidence of a given cell c.

- G_c is the ground truth of a given cell c.

$$\text{IoU}(P, G) = \frac{\sum |P_c \cap G_c|}{\sum |P_c \cup G_c|} \tag{3.6}$$

A helpful visualisation of how this will work in our study is illustrated in Figure 3.18.

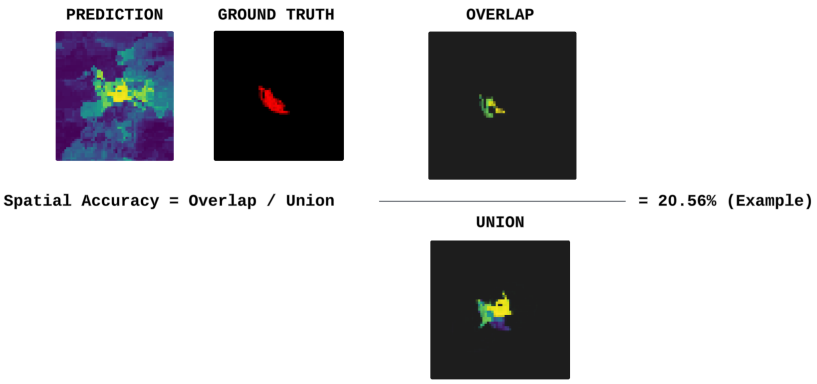

Figure 3.18: Visualisation of IoU.

3.5.6 Hold-Out Validation

We will primarily be investigating model performance and fit using Hold-Out, this validation technique involves fragmenting the dataset into train/validation/test splits whereby all sets contains unique data. After model training analysis, we will enact full retraining where we concatenate training and validation to narrow the split to just training/test.

3.5.7 K-Fold Cross-Validation

Though Hold-Out is the better validation strategy, we also need to employ K-Fold Cross-Validation as it involves distributing the the dataset into *k* folds, where training is done on the other folds and evaluation on a *k-1* fold. This ensures that all the data can be utilised without learning on data it is predicting. We require this technique not only to address smaller datasets, which we will encounter with true fossil occurrences, but also to produce a complete map of predictions for the spatial model.

3.5.8 Hyperparameter Tuning

Hyperparameter tuning can be executed through several strategies: regularisation optimisation, capacity optimisation, and broader model refinement. A range of both simple and advanced algorithms are available in the Keras Library. Traditionally, Grid Search is employed to systematically explore Hyperparameter combinations while balancing resources efficiency. However, we will leverage Hyperband Optimisation, which iteratively navigates the search space rather than exhausting every permutation. This approach allows for a more efficient and targeted search, honing in on the optimal values with fewer iterations, but, our search depth is subjective and can lead to extremely long tuning times, hence we will focus on regularisation, minimal layer capacity improvements and broader refinement.

3.5.9 Research Questions

Now that we have discussed methodology, we can establish our research questions. Our two primary hypotheses are:

1. *"By leveraging temporal patterns in environmental and geological data, the model can predict fossiliferous regions with high accuracy. The model's predictions will sig-*

nificantly outperform random guessing and demonstrate feasibility over non-temporal methods."

2. *"The post-processed suitability result will accurately represent fossiliferous regions, rarely missing any true positives. The system will generalise to predict unexplored regions, and spatial accuracy will be consistent with ground-truth fossil patterns."*

These hypotheses outline success in the study's approach, we can further our experimental analysis by attempting to answer two more questions:

1. *"Will changing temporal steps impact the model's predictive power?"*

2. *"Have we elected meaningful feature extractions and what was their influence on the models?"*

For this study we can quantify the degree of success by establishing a series of goals and determine metric-specific thresholds to justify success in various aspects of performance.

Goal	**Metric Threshold**
Ensure both models can beat baseline metrics	Greater than random guessing figures.
Ensure both models achieve validation metrics	AUC-ROC $\geqq$ 0.85%. Precision / Recall $\geqq$ 0.95%.
Ensure the CNN model achieves persuasive spatial metrics	Spatial extent of $\geqq$ 0.30%.
Harmonic Suitability Result	Identify $\geqq$ 0.90% of known fossil locations with high-confidence.
Hypothesis 1 Evaluation	AUC-ROC $\geqq$ 0.80%. Precision / Recall $\geqq$ 0.90%. IoU $\geqq$ 0.20%.
Hypothesis 2 Evaluation	AUC-ROC $\approx$ 10-20% MoE. Precision / Recall $\approx$ 10-20% MoE. IoU $\geqq$ 0.50%.

Table 3.7: Goals and corresponding metric-thresholds.

Chapter 4

Implementation and Results

This section outlines our experiment ideas, model workflows, preliminary and post-result analysis.

4.1 Comparative Setup

In order to gauge success, we must first establish a form of comparative measure. We will collate metrics from both random guessing and several logistic regressions, additionally we will perform Feature-Importances to verify suitability of our features.

4.1.1 Baselines

Model	**Feature Input**	**Reasoning**
Logistic T(1)	Only T1 Feature Data	Whether a non-temporal approach is better for this type of problem.
Logistic Flattened	T1 $\rightarrow$ T50 Data as 1 Feature	Whether the temporal data is independent, i.e., does temporal order matter?
Logistic Mean Average	T1 $\rightarrow$ T50 Data averaged into a T(1) Feature	Whether we have a general trend across all of time. Does long-term data integrity matter instead of short-term?

Table 4.1: Comparison of Different Logistic Models.

We can also deduce preliminary metric conclusions if we are to prove our hypotheses correct:

Model	Expectation
Logistic T(1)	Demonstrate good predictive power if T2 → T50 are repeating information or irrelevant, suggesting sequential data is not required for the classification task.
Logistic Flattened	Demonstrate the best metrics if sequential data is important without ordinal relationships.
Logistic Mean Average	Demonstrate the best metrics if T1 → T50 share similar geologic and environmental representations.

Table 4.2: Assumptions of Logistic Models.

4.1.2 Feature Importances

To understand the impact of each feature on predictions we need to run an implementation of feature importance. We chose to implement a Random Forest Classifier (Listing 4.1).

```
rf = RandomForestClassifier(n_estimators=100, random_state=42, n_jobs=-1)
rf.fit(features, labels)
```

Listing 4.1: Random Forest at 100 estimators (Github).

We can then construct inferences on both features across all time-steps and features at each time-step as shown in Listing 4.2.

```
# Aggregated Feature Importance (by each feature)
featureImportances["BaseFeature"] = featureImportances["Feature"].str.split('_T').str[0]
baseImportances = featureImportances.groupby("BaseFeature")["Importance"].sum().reset_index()
baseImportances = baseImportances.sort_values(by="Importance", ascending=False)

# Aggregated Feature Importance (by timestep)
featureImportances["Timestep"] = featureImportances["Feature"].str.split('_T').str[1].astype(float)
timestepImportances = featureImportances.groupby("Timestep")["Importance"].sum().reset_index()
timestepImportances = timestepImportances.sort_values(by="Timestep", ascending=True)
```

Listing 4.2: Sorting importances by feature and by time-step (Github).

4.2 Temporal Viability

4.2.1 Sedimentary vs. Non-Sedimentary

We can begin our first experiment by investigating if the model and features are able to hone in on sedimentary lithologies. We will not spend much time in exploiting the best model and metrics but focus on feasibility. The following describes our experimental setup:

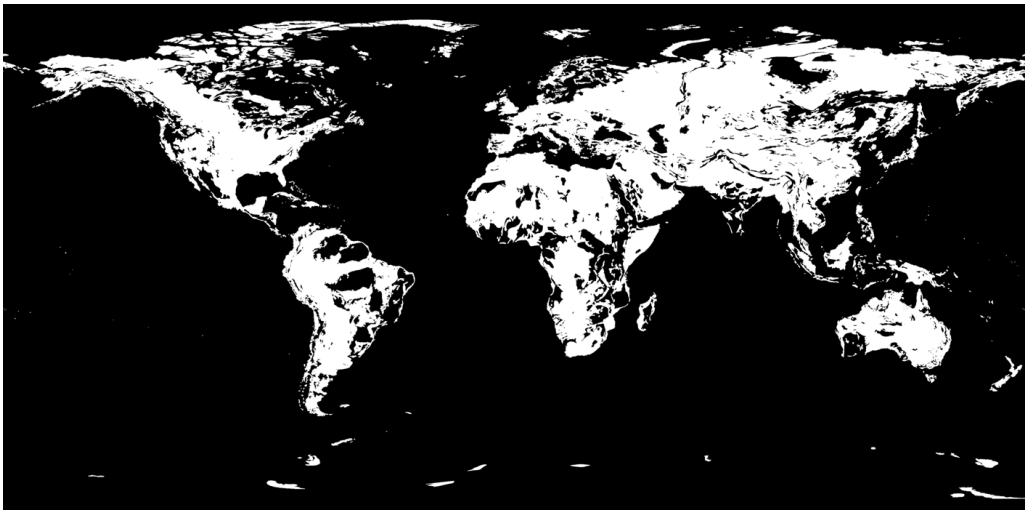

Figure 4.1: Raster of Target Label (Black = Non-Sedimentary/Water, White = Sedimentary Landmass).

Features	**Labels(s)**
Precipitation	Sedimentary/Non-Sedimentary Map (Figure 4.1)
Elevation	-
Temperature	-
Water Flux	-
Sediment Flux	-
Flood Basins (Lakes)	-
Erosion Rate	-
Uplift Rate	-
Slope	-
Flow Accumulation	-
Death Signal (Death-Densities)	-

Table 4.3: Experimental Dataset.

One important discrepancy to note is that our target infers that we are sampling the entire raster, this is not true - we sample only landmass cells.

Unregularised Model

Our baseline RNN model starts with a simple dual stacked LSTM followed by a dense block; we also apply batch normalisation to aid learning given the numerous permutations every sample may introduce.

Layer	Output Shape	Parameters
Input	(None, 50, 11)	-
LSTM 1	(None, 50, 256)	274,432
Batch Normalization	(None, 50, 256)	1,024
LSTM 2	(None, 160)	266,880
Batch Normalization	(None, 160)	640
Fully Connected 1 (ReLU)	(None, 128)	20,608
Fully Connected 2 (Sigmoid)	(None, 1)	129
	Total	**563,713**

Table 4.4: Dual-Stack LSTM.

Test Set	Binary Threshold	Metric	Result	Baseline	Beaten?
[13157, 30897]	0.90	Accuracy	0.8054	0.5797	Yes
	”	Precision	0.9909	0.7028	Yes
	”	Recall	0.7292	0.7000	No
	”	F1-Score	0.8401	0.7014	Yes
	N/A	Confidence	0.9239	0.1816	Yes
	0.90	AUC-ROC	0.9606	0.50	Yes
	”	Theoretical Baseline	0.7013	N/A	Yes

Table 4.5: (Model 4.4) Baseline Metrics Comparison.

Upon acceptable results, we then compare performance to the three logistic models:

Metric	**Binary Threshold**	**Logit T(1)**	**Logit Flat**	**Logit Avg**	**LSTM Baseline**	**Beaten?**
Accuracy	0.90	0.3054	0.3906	0.3001	0.8054	No
Precision	”	0.9512	0.9692	0.8605	0.9909	No
Recall	”	0.0101	0.1354	0.0024	0.7292	No
F1-Score	”	0.0200	0.2376	0.0048	0.8401	No
Confidence	N/A	0.7340	0.8170	0.7203	0.9239	No
AUC-ROC	0.90	0.6677	0.8233	0.6288	0.9606	No

Table 4.6: Metrics Comparisons Between Logistic and LSTM.

Table 4.6 evinces our presumptions about temporal behaviour. We can see that the simplicity of logistic models prevent a deeper understanding of classification, additionally the best performing logistic model was *Logit Flat* which disregards ordinance, proving temporal feasibility.

Regularised Model

Now that we have proven statistical power, we can address overfitting through regularisation techniques - we apply dropouts layers after each block.

Layer	**Output Shape**	**Parameters**
Input	(None, 50, 11)	-
LSTM 1	(None, 50, 256)	274,432
Batch Normalization	(None, 50, 256)	1,024
Dropout	(None, 50, 256)	0
LSTM 2	(None, 160)	266,880
Batch Normalization	(None, 160)	640
Dropout	(None, 160)	0
Fully Connected 1 (ReLU)	(None, 128)	20,608
Dropout	(None, 128)	0
Fully Connected 2 (Sigmoid)	(None, 1)	129
	Total	**563,713**

Table 4.7: Dual-Stack LSTM with Light Regularisation.

To figure out whether or not regularisation is effective, we need to inspect training graphs, such that we can follow loss and accuracy trends to determine whether we are still overfitting,

underfitting or found a robust fit. Figure 4.2 shows the graphs for the regularised model.

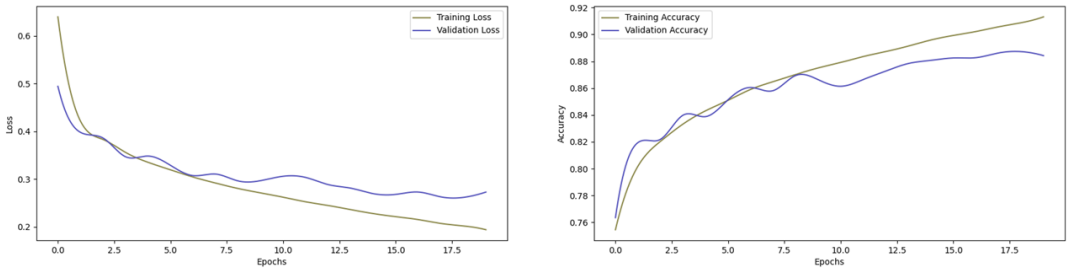

Figure 4.2: (Model 4.7) Training Graphs. Validation (Purple) represents model performance on unseen data, if the model cannot train and validate at a consistent rate there is an issue to do with the model setup approach.

We observe overfitting towards the end of the training epoch range, it exhibits instability in training and quite early divergence in loss, an optimal epoch at approximately 15 suggests that the quick dropouts were of some benefit but not entirely, we can potentially say that even with the dropout rates the model could do with an in increase in capacity and further regularisation.

Test Set	**Binary Threshold**	**Metric**	**Reg LSTM**	**Unreg LSTM**	**Beaten?**
[13157, 30897]	0.90	Accuracy	0.8131	0.8054	Yes
	”	Precision	0.9894	0.9909	No
	”	Recall	0.7415	0.7292	Yes
	”	F1-Score	0.8477	0.8401	Yes
	N/A	Confidence	0.9193	0.9239	No
	0.90	AUC-ROC	0.9616	0.9606	Yes

Table 4.8: (Model 4.7) Regularised vs. (Model 4.4) Unregularised Comparison.

We also observe a marginal improvement over the unregularised model, this is due to lack of aggressiveness from regularisation.

Tuned Model

With an increase to capacity, Hyperband Optimisation was employed to tune dropouts and dense layer kernel-regularisers individually to get the an appropriate permutation:

Layer	Output Shape	Parameters
Input	(None, 50, 11)	-
LSTM 1	(None, 50, 480)	944,640
Batch Normalization	(None, 50, 480)	1,920
Dropout	(None, 50, 480)	0
LSTM 2	(None, 256)	754,688
Batch Normalization	(None, 256)	1,024
Dropout	(None, 256)	0
Fully Connected 1 (ReLU)	(None, 128)	32,896
Dropout	(None, 128)	0
Fully Connected 2 (Sigmoid)	(None, 1)	129
	Total	**1,735,297**

Table 4.9: Higher Capacity Dual-Stack LSTM.

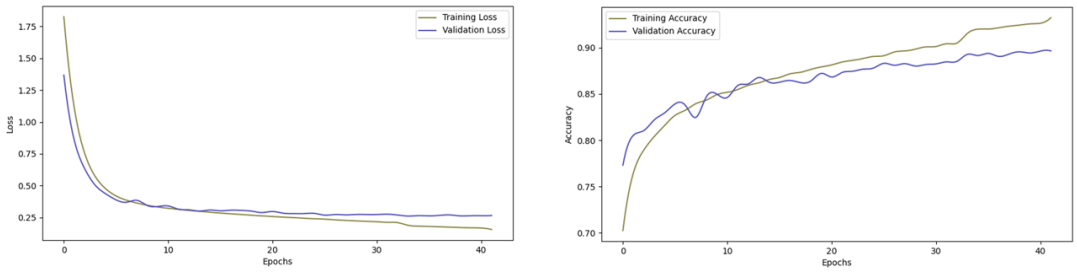

Figure 4.3: (Model 4.9) Training Graphs.

The training graphs show a significant improvement to model fit, we can see that up until Epoch 37 there is not divergence in loss. We then do a complete retraining at that optimal Epoch to deduce our tuned results.

Test Set	Binary Threshold	Metric	Tuned LSTM	Untuned LSTM	Beaten?
[13157, 30897]	0.90	Accuracy	0.8445	0.8131	Yes
	”	Precision	0.9848	0.9894	No
	”	Recall	0.7905	0.7415	Yes
	”	F1-Score	0.8770	0.8477	Yes
	N/A	Confidence	0.9340	0.9193	Yes
	0.90	AUC-ROC	0.9643	0.9616	Yes

Table 4.10: (Model 4.9) Tuned vs. (Model 4.7) Untuned Comparison.

Furthermore, we attempted to compare LSTM vs. GRU to validate any claims that recent temporal changes are indeed more significant than long-term changes:

Test Set	Binary Threshold	Metric	GRU	LSTM	Beaten?
[13157, 30897]	0.90	Accuracy	0.8398	0.8445	No
	”	Precision	0.9855	0.9848	Yes
	”	Recall	0.7832	0.7905	No
	”	F1-Score	0.8727	0.8770	No
	N/A	Confidence	0.9363	0.9340	Yes
	0.90	AUC-ROC	0.9643	0.9643	No

Table 4.11: LSTM vs. GRU Comparison.

4.2.2 Results

Results confirm that temporal modelling enhances preservation predictions over alternative methods. *Why does temporality improve results?* Likely due to the gradual/evolutional nature of sedimentary deposition, further evidenced by the weakened predictive power, as shown in Figure 4.4.

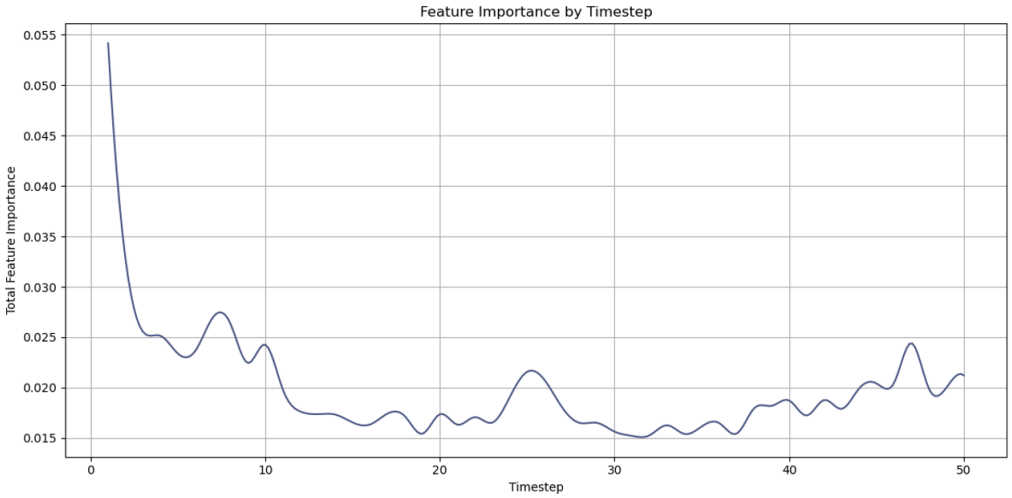

Figure 4.4: Importance by Time-step. Non-linearity is observed, with higher importance closer to T(1). Temporal integrity is further supported by single time-step logit models where historical data is absent.

Which features contributed most? Figure 4.5 highlights a weak feature, 'Death Signal,' suggesting either an insufficient temporal window or geological inconsistency over time. Other features align with expected importance rankings.

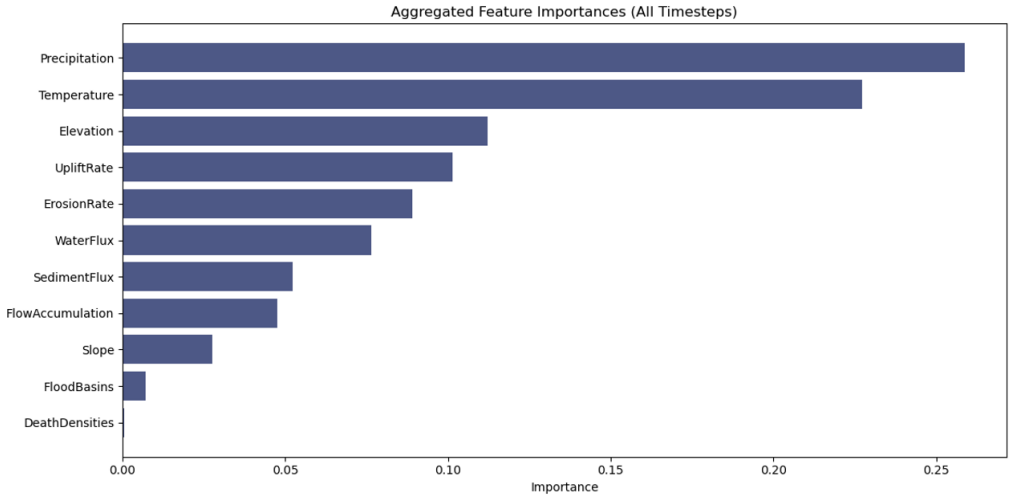

Figure 4.5: Importance by Features. Environmental factors display expected significance.

GRU and LSTM models perform nearly identically (4.11), suggesting long-term dependencies are less critical than anticipated. Nonetheless, RNNs remain the superior choice for this type of prediction.

4.2.3 Time-step Significance

In this experiment, we assess whether a 5 Myr interval provides sufficient temporal resolution and if coarser intervals degrade predictive power. To evaluate this, we can investigate three intervals against our logistic suite.

Metric	**Binary Threshold**	**Logit T(1)**	**Logit Flat**	**Logit Avg**
Accuracy (5 Myr)	0.90	0.3054	0.3906	0.3001
Accuracy (10 Myr)	0.90	0.2987	0.3314	0.2987
Accuracy (25 Myr)	0.90	0.2987	0.3090	0.2988

Metric	**Binary Threshold**	**Logit T(1)**	**Logit Flat**	**Logit Avg**
Precision (5 Myr)	0.90	0.9512	0.9692	0.8605
Precision (10 Myr)	0.90	0.0000	0.9327	1.0000
Precision (25 Myr)	0.90	0.0000	0.9236	1.0000

Metric	**Binary Threshold**	**Logit T(1)**	**Logit Flat**	**Logit Avg**
Recall (5 Myr)	0.90	0.0101	0.1354	0.0024
Recall (10 Myr)	0.90	0.0000	0.0503	0.0000
Recall (25 Myr)	0.90	0.0000	0.0161	0.0002

Metric	**Binary Threshold**	**Logit T(1)**	**Logit Flat**	**Logit Avg**
F1-Score (5 Myr)	0.90	0.0200	0.2376	0.0048
F1-Score (10 Myr)	0.90	0.0000	0.0954	0.0001
F1-Score (25 Myr)	0.90	0.0000	0.0316	0.0004

Metric	Binary Threshold	Logit T(1)	Logit Flat	Logit Avg
Confidence (5 Myr)	0.90	0.7340	0.8170	0.7203
Confidence (10 Myr)	0.90	0.7016	0.7912	0.7094
Confidence (25 Myr)	0.90	0.7054	0.7672	0.7072

Metric	Binary Threshold	Logit T(1)	Logit Flat	Logit Avg
AUC-ROC (5 Myr)	0.90	0.6677	0.8233	0.6288
AUC-ROC (10 Myr)	0.90	0.5022	0.7830	0.5928
AUC-ROC (25 Myr)	0.90	0.5546	0.7391	0.5738

4.2.4 Results

Metrics confirm that predictive power declines as time-step intervals widen, supporting findings from 4.2.2.

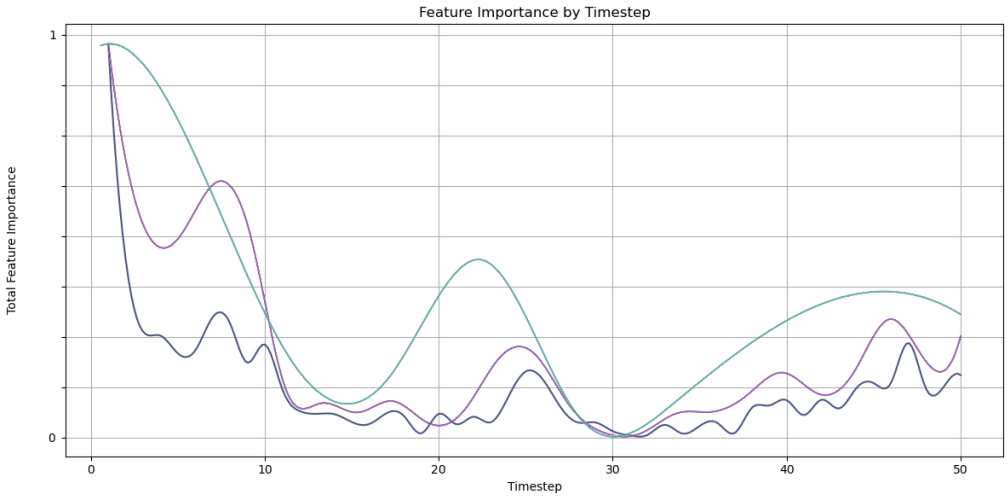

Figure 4.6: Visualisation of time-step importances (Dark Blue = 5 Myr, Purple = 10 Myr, Turquoise = 25 Myr).

Figure 4.6 demonstrates the loss of temporal fidelity. While larger intervals retain some patterns, they become too coarse for further deductions. The key question remains: *Are 5 Myr increments sufficient?* - for this study, yes. Across all models, we avoid extreme metric collapse which would occur at 10+ Myr intervals.

4.3 Geologic Period Preservation Model

This stage focuses on categorising regions of sedimentary land into their respective temporal ages, to simulate richly sampled domains without absence uncertainty, before finally electing the appropriate model to base our Transfer-learning.

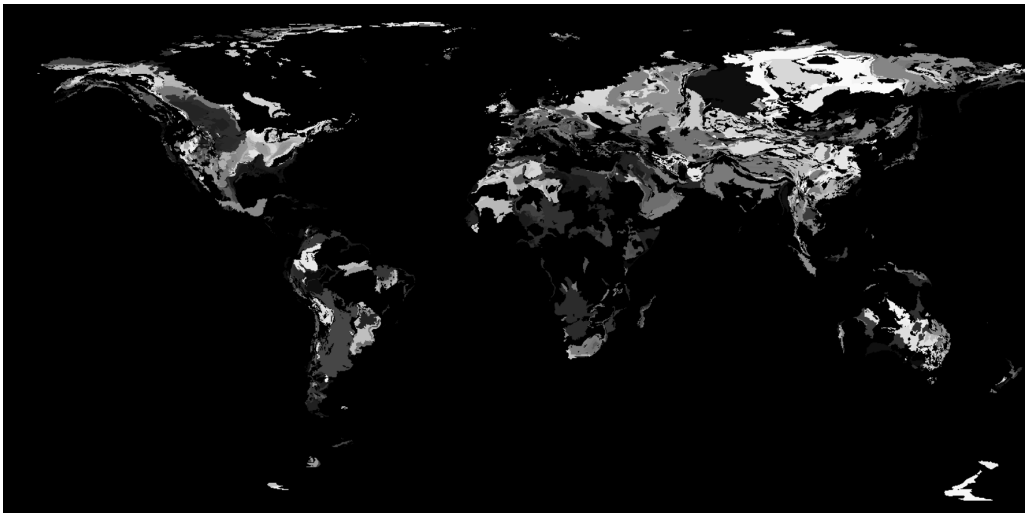

Figure 4.7: Raster of Target Label (Black = Non-Sedimentary/Water/Unknown, Darker = Younger, Lighter = Older).

Features	**Labels(s)**
Precipitation	Aged Sedimentary Map (Categorised) (Figure 4.7)
Elevation	-
Temperature	-
Water Flux	-
Sediment Flux	-
Flood Basins (Lakes)	-
Erosion Rate	-
Uplift Rate	-
Slope	-
Flow Accumulation	-

Table 4.12: Experimental Dataset.

Due to per-polygon age-ing, we had 2,752 discrete values which needed attenuation to make the data eligible for multi-class prediction. We used one-hot-encoding:

Encoded Category	Range
Pre-Mesozoic	$\geqq 250$
Triassic	$\geqq 199$
Jurassic	$\geqq 141$
Cretaceous	$\geqq 64$
Cenozoic	$\geqq 21$
N/A	< 21

Table 4.13: Encoded Ranges with ± 2 Myr Buffer.

4.3.1 Single-Head RNNs

Unregularised Model

The model capacity from the previous stage was inferred as fundamentally the features and temporal depth are relatively the same. The innate complexity that a 6-class classifier will most likely suggest that the model will be at capacity, or in some sense, under/overfit.

Layer	Output Shape	Parameters
Input	(None, 50, 10)	-
LSTM 1	(None, 50, 480)	942,720
Batch Normalization	(None, 50, 480)	1,920
LSTM 2	(None, 256)	754,688
Batch Normalization	(None, 256)	1,024
Fully Connected 1 (ReLU)	(None, 128)	32,896
Fully Connected 2 (Softmax)	(None, 6)	774
	Total	**1,734,022**

Table 4.14: Dual-Stack LSTM with Softmax Classifier.

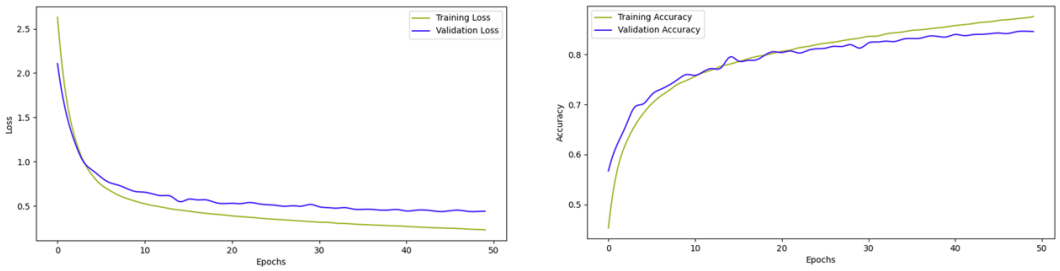

Figure 4.8: (Model 4.14) Training Graphs.

Training graphs showed that the model is at an appropriate capacity as loss and accuracies stay close, no divergence or large gaps across the training. We observe acceptable

performance across all classes as shown in Figure 4.9.

Encoded Category	Precision	Recall	F1-Score	Support
N/A	0.9471	0.8013	0.8681	20392
Cenozoic	0.7602	0.9281	0.8358	2378
Cretaceous	0.7483	0.9122	0.8221	4337
Jurassic	0.6777	0.9000	0.7732	2311
Triassic	0.6334	0.8779	0.7359	1655
Pre-Mesozoic	0.8490	0.8666	0.8577	12981

Table 4.15: (Model 4.14) per-class Metrics.

Table 4.15 shows performance decline for Triassic/Jurassic with noticeably lower Precision and Recall compared to younger periods. Despite Jurassic having a sample size comparable to Cenozoic, metrics suggest that the model is exhibiting difficulty in feature extraction. This could be due to temporal compression with just two LSTM layers.

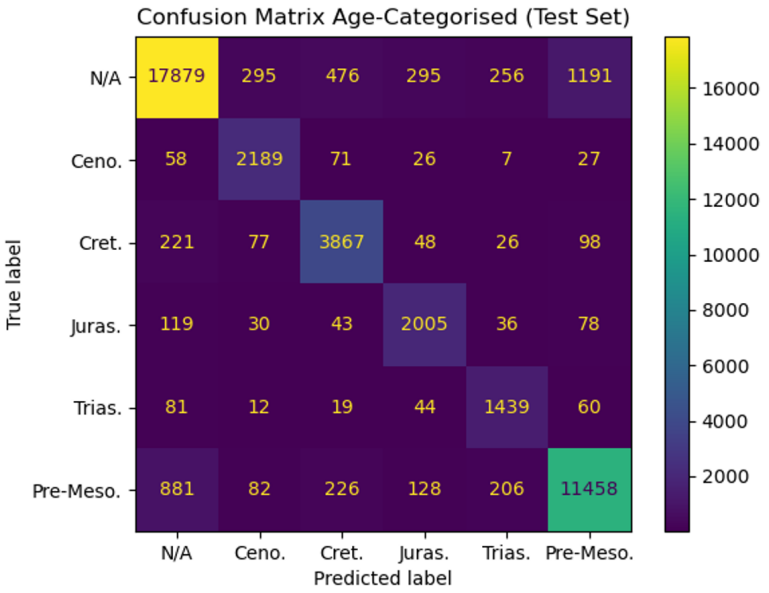

Figure 4.9: (Model 4.14) Confusion Matrix.

Comparison Models

We observe closer delta metric performances in Logistic Models compared to the previous stage (4.2).

Metric	**Binary Threshold**	**Logit T(1)**	**Logit Flat**	**Logit Avg**	**LSTM Baseline**	**Beaten?**
Accuracy	N/A	0.2762	0.5820	0.2888	0.8463	No
Precision	”	0.2557	0.4975	0.2675	0.7693	No
Recall	”	0.3122	0.6646	0.3283	0.8810	No
F1-Score	”	0.2326	0.5358	0.2485	0.8155	No
AUC-ROC	”	0.6618	0.8877	0.6777	0.9821	No

Table 4.16: Metrics Comparison Between Logistic and LSTM Models.

Test Set	**Binary Threshold**	**Metric**	**GRU**	**LSTM**	**Beaten?**
[20392 2378 4337 2311 1655 12981]	N/A	Accuracy	0.8393	0.8463	No
	”	Precision	0.7378	0.7693	No
	”	Recall	0.8708	0.8810	No
	”	F1-Score	0.7901	0.8155	No
	”	AUC-ROC	0.9802	0.9821	No

Table 4.17: LSTM vs. GRU Comparison.

Segregating sediment regions into temporal confinements does not enhance distinctions between long-term and short-term changes, yet neither does it change the results between GRU vs. LSTM (4.17) - complimenting the findings of the previous binary classification.

The model tends to misclassify *N/A* which represents non-sedimentary/unknown geologic age, this error can be attributed to spatial inaccuracies within dataset polygons due to resolution and data accuracy, as well as under-sampling in regions where there is not enough occurrence information to derive conclusive age estimates.

Tuned Model

As we already proved good training, the model doesn't need strenuous dropouts, therefore we can increase the capacity further to allow for some space for any regularisation and encourage further overfitting to ensure we have the best model. Hyperband tuning resulted in a slightly higher capacity setup:

Layer	**Output Shape**	**Parameters**
Input	(None, 50, 10)	-
LSTM 1	(None, 50, 512)	1,071,104
Batch Normalization	(None, 50, 512)	2,048
Dropout	(None, 50, 512)	0
LSTM 2	(None, 256)	922,752
Batch Normalization	(None, 256)	1,152
Dropout	(None, 256)	0
Fully Connected 1 (ReLU)	(None, 128)	18,496
Dropout	(None, 64)	0
Fully Connected 2 (Softmax)	(None, 6)	390
	Total	**2,015,942**

Table 4.18: Tuned Dual-Stack LSTM with Softmax Classifier.

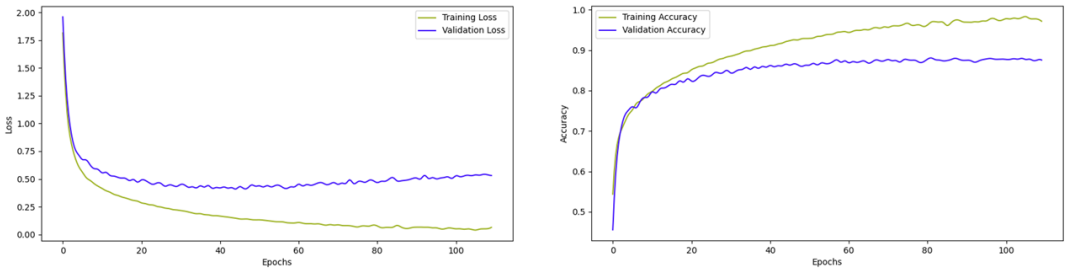

Figure 4.10: (Model 4.18) Training Graphs.

At a full retraining we were able to improve the Jurassic/Triassic performances by a large margin.

Encoded Category	Precision	Recall	F1-Score	Support
N/A	0.9293	0.8768	0.9023	20392
Cenozoic	0.8153	0.9205	0.8647	2378
Cretaceous	0.8224	0.8916	0.8556	4337
Jurassic	0.7875	0.8676	0.8256	2311
Triassic	0.7305	0.8695	0.7939	1655
Pre-Mesozoic	0.8874	0.8827	0.8850	12981

Table 4.19: (Model 4.18) per-class Metrics.

Now that we've concluded the best training and Hyperparameter search we can experiment further by taking a different approach to feature inputs.

4.3.2 Multi-Head LSTM

This experiment's goal is to see if our LSTM layers are being oversaturated with the number of features and if we can improve predictive power. The new architecture will train several heads on different features, apply an attention mechanism and therefore give bias towards stronger neurons on a smaller group of features. We adopted a three-group split, where features fall into: Climate, Hydro, Erosion. We grouped features as shown in Table 4.20.

Climate	Hydro	Erosion
Precipitation	Flow Accumulation	Erosion Rate
Elevation	Water Flux	Uplift Rate
Temperature	Sediment Flux	Slope
-	Flood Basins (Lakes)	-

Table 4.20: Features Grouped.

We chose to retain the dual-stack layers but incorporate bi-directional support. We did not anticipate any substantial improvements but rather attempt to address outliers. A regularised implementation of the model is as follows:

Layer	Output Shape	Parameters
Input Climate	(None, 50, 3)	-
Input Hydro	(None, 50, 4)	-
Input Erosion	(None, 50, 3)	-

Table 4.21: Split Input Layers.

Layer	Output Shape	Parameters
Bidirec. LSTM Climate 1	(None, 50, 512)	532,480
Bidirec. LSTM Hydro 1	(None, 50, 640)	832,000
Bidirec. LSTM Erosion 1	(None, 50, 512)	532,480
Batch Normalization Climate	(None, 50, 512)	2,048
Batch Normalization Hydro	(None, 50, 640)	2,560
Batch Normalization Erosion	(None, 50, 512)	2,048
Dropout Climate	(None, 50, 512)	0
Dropout Hydro	(None, 50, 640)	0
Dropout Erosion	(None, 50, 512)	0
Bidirec. LSTM Climate 2	(None, 50, 192)	467,712
Bidirec. LSTM Hydro 2	(None, 50, 320)	1,025,280
Bidirec. LSTM Erosion 2	(None, 50, 192)	467,712
Batch Normalization Climate	(None, 50, 192)	768
Batch Normalization Hydro	(None, 50, 320)	1,280
Batch Normalization Erosion	(None, 50, 192)	768
Dropout Climate	(None, 50, 512)	0
Dropout Hydro	(None, 50, 640)	0
Dropout Erosion	(None, 50, 512)	0

Table 4.22: LSTM Blocks with Normalisation and Dropouts.

Layer	Output Shape	Parameters
Multi-Head Attention Climate	(None, 50, 192)	98,880
Multi-Head Attention Hydro	(None, 50, 320)	164,544
Multi-Head Attention Erosion	(None, 50, 192)	98,880
Residual Connection Climate	(None, 50, 192)	0
Residual Connection Hydro	(None, 50, 320)	0
Residual Connection Erosion	(None, 50, 192)	0
Layer Normalization Climate	(None, 50, 192)	384
Layer Normalization Hydro	(None, 50, 320)	640
Layer Normalization Erosion	(None, 50, 192)	384
Global Pooling 1D Climate	(None, 50, 192)	0
Global Pooling 1D Hydro	(None, 50, 320)	0
Global Pooling 1D Erosion	(None, 50, 192)	0

Table 4.23: Multi-Head and Normalisation.

Layer	Output Shape	Parameters
Concatenate Heads	(None, 704)	0
Batch Normalization	(None, 704)	2,816
Dropout	(None, 704)	0
Fully Connected 1 (ReLU)	(None, 128)	90,240
Batch Normalization	(None, 128)	512
Dropout	(None, 128)	0
Fully Connected 2 (ReLU)	(None, 64)	8,256
Dropout	(None, 64)	0
Fully Connected 3 (Softmax)	(None, 6)	390
	Total	**4,333,062**

Table 4.24: Final Dense Layers with Softmax Classifier.

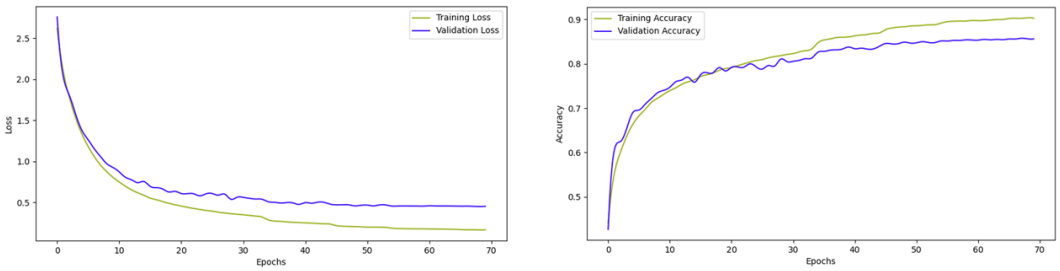

Figure 4.11: (Model 4.24) Training Graphs.

We found a variety of classification improvements, specifically Recall.

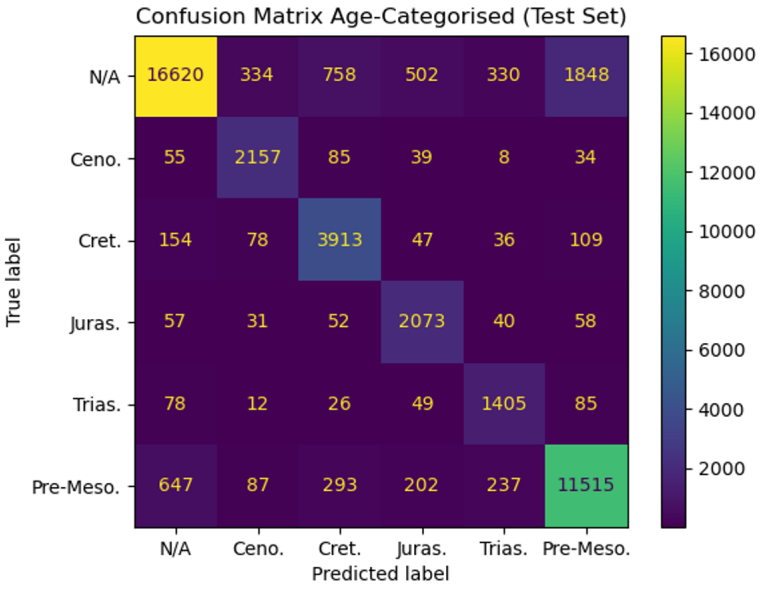

Figure 4.12: (Model 4.24) Confusion Matrix.

After subsequent regularisation tuning and full retraining we further increased our performance.

Encoded Category	Precision	Recall	F1-Score	Support
N/A	0.9411	0.8312	0.8828	20392
Cenozoic	0.8223	0.9108	0.8643	2378
Cretaceous	0.7960	0.8995	0.8446	4337
Jurassic	0.7349	0.9022	0.8100	2311
Triassic	0.7047	0.8653	0.7768	1655
Pre-Mesozoic	0.8485	0.8915	0.8695	12981

Table 4.25: (Model 4.24) Tuned Metrics.

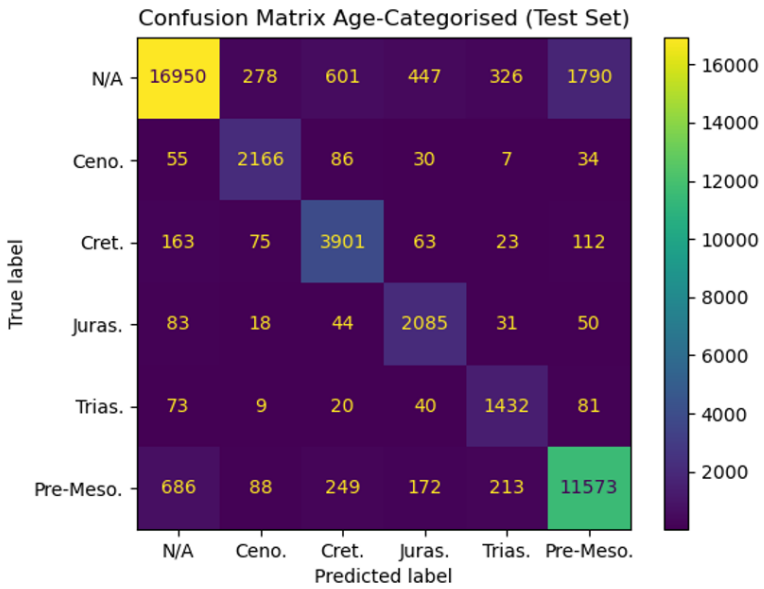

Figure 4.13: Observed improvements in comparison to the untuned model (Figure 4.12).

4.3.3 Results

We found that the previous target and this multi-class target shared frequent similarities in terms of model appropriation and importances.

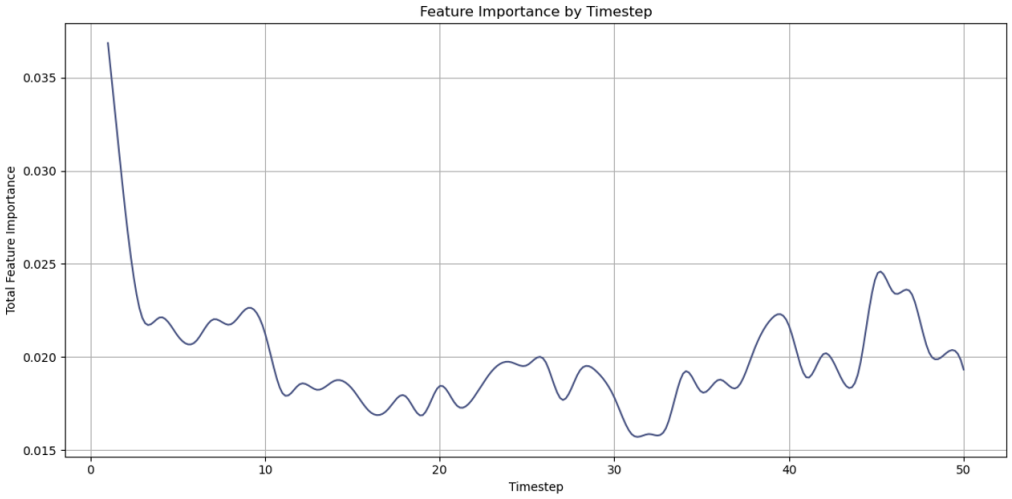

Figure 4.14: Importance by Time-step.

Figure 4.14 demonstrates same parabolic trend as Figure 4.4.

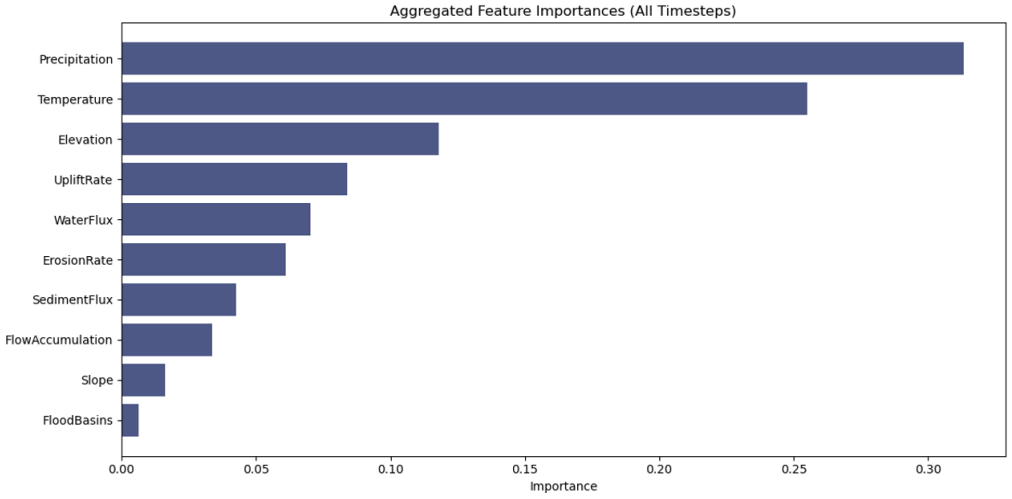

Figure 4.15: Importance by Features.

In regards to our chosen model, there are nominal differences between both single/Multi-Head models, but the Multi-Head Confusion Matrix (Figure 4.13) illustrates improvements to all Mesozoic-scope categories. Ultimately, both models are strong and perform almost identically with our assumption of feature confusion being proved wrong but not to the extent of dismissing hidden nuanced signals that would justify the improved Recall.

Test Set	**Binary Threshold**	**Metric**	**Multi-Head**	**Regular**
[20392 2378 4337 2311 1655 12981]	N/A	Accuracy	0.8650	0.8816
	”	Precision	0.8079	0.8287
	”	Recall	0.8834	0.8848
	”	F1-Score	0.8413	0.8545
	”	AUC-ROC	0.9853	0.9863

Table 4.26: Multi-Head vs. Non-Multi-Head Comparison.

Which model will we proceed with? To which we will choose the Multi-Head as we wish to utilise the marginal recall-boost presented to enhance positive class encounters in our deeper-time categories.

4.4 Transfer-learned Preservation Model

To conclude our temporal aspect of the study, we will adapt our previous model to focus on both true fossiliferous occurrences and geologically specific regions.

4.4.1 Transfer-learned RNNs

Our experimental setup:

Figure 4.16: Raster of Holistic Target (Black = Non-Occurrence, White = Recorded-Occurrence).

Features	Labels(s)
All Section 4.3 Features	Cretaceous/Jurassic/Triassic/Holistic Occurrences (e.g., Figure 4.16)

Table 4.27: Experimental Dataset.

Our solution to address the problem of finding appropriate pseudo absences is attempted solution by randomly sampling cells (Listing 4.3) that have had no fossil occurrences relative to the geologic period of focus, softly suggesting true negatives.

```
# Get indices for where we know there has been fossil activity
trueSamples = np.where(cellLabels == 1)[0]
# Get all the other indices where there's not been any reported fossil
    activity relative to the geologic period
negativeSamples = np.where(cellLabels == 0)[0]
# Randomly choose from the negative group to satiate the same true
    negatives for the model on every run
```

```
np.random.seed(42)
# Currently trying not to oversample negatives to prevent the model from catering against the true positives
randomNegativeSampledIndices = np.random.choice(negativeSamples, min(2 * len(trueSamples), len(negativeSamples)), replace=False)
# Combine true and false
allSamplesIndices = np.concatenate([trueSamples, randomNegativeSampledIndices])
# Update our labels and features with the new subset
cellLabels = cellLabels[allSamplesIndices]
cellFeatures = cellFeatures[allSamplesIndices]
```

Listing 4.3: Randomly Sampling Negatives (Github).

Although this should help the model, there are some underlaying limitations and issues. Nevertheless it should suffice to determine the feasibility of our research aims and hypotheses.

Comparison Models

We will reiterate arduousness by presenting the logistic performances of this experiment, we split the dataset into a 2:1 ratio for holistic occurrences. We elected this size to both keep conservative and not undersample our positive samples. Models showed that they were either just at statistical significance or marginally above:

Metric	**Binary Threshold**	**Logit T(1)**	**Logit Flat**	**Logit Avg**
Accuracy	0.90	0.6666	0.6971	0.6658
Precision	”	0.0000	0.8082	0.3947
Recall	”	0.0000	0.1198	0.0045
F1-Score	”	0.0000	0.2087	0.0090
Confidence	N/A	0.3586	0.4799	0.3708
AUC-ROC	0.90	0.6344	0.8090	0.6842

Table 4.28: Metric collapse in T(1) and T(Average).

Candidate Model

To conduct Transfer-learning, we will take the existing Model (4.24), freeze all layers and replace the last Dense-Block with a couple layers, ReLU-activated and Sigmoid-activated.

Layer	Output Shape	Parameters
Concatenate Heads	(None, 704)	0
Batch Normalization	(None, 704)	2,816
Dropout	(None, 704)	0
Fully Connected 1 (ReLU)	(None, 64)	45,120
Fully Connected 3 (Sigmoid)	(None, 1)	65
	Total	**4,278,849**

Table 4.29: New Dense-Block with Sigmoid Classifier.

A subsequent Hold-Out training run showed that we had a correct capacity as the model retained stable training but exhibited overfitting.

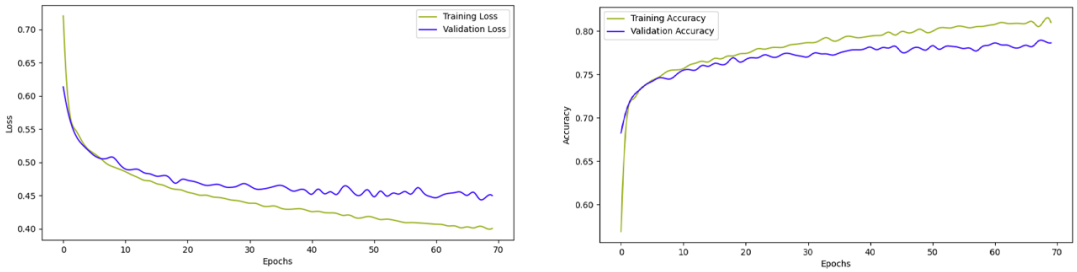

Figure 4.17: (Model 4.29) demonstrating overfitting.

Tuned Model

Once tuned, the model trained very effectively, overfitting at a high Epoch of approximately 85.

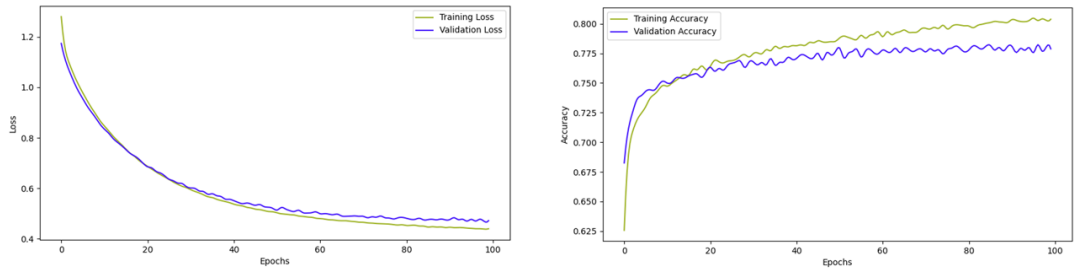

Figure 4.18: (Model 4.29) showing sapid learning across the training.

4.4.2 Results

In order to get the best generalisation performance, we alternated to K-Fold CV and evaluated Mesozoic-scope categories individually to confirm consistency and reliability in the model.

Binary Threshold	**Metric**	**All Fossils**	**Cretaceous**	**Jurassic**	**Triassic**
0.90	Accuracy	0.7277	0.8267	0.8530	0.8393
”	Precision	0.8815	0.8313	0.8569	0.8471
”	Recall	0.2115	0.3848	0.4947	0.4359
”	F1-Score	0.3412	0.5261	0.6273	0.5756
N/A	Confidence	0.5495	0.5207	0.5533	0.5441
0.90	AUC-ROC	0.8697	0.9092	0.9249	0.9230
”	IoU	0.1934	0.3569	0.4569	0.4041
	Dataset Ratio	2:1	3:1	3:1	3:1

Table 4.30: K(5)-Fold Results.

One thing to note, we observe higher performances from the isolated categories but this is a result of an increase to the class imbalance, instead of dealing with $\frac{2}{3}$% negatives, we've increased it to $\frac{3}{4}$% to amplify the dataset size. Regardless, the model still perform very well in most metrics. Despite emphasis for a higher Precision, we've discovered a subpar Recall/IoU, lower than 50% suggests that the model cannot predict most known sites reliably. We know that this is not threshold-related as our Confidence scores are similar, hence this is an modelling/data issue.

Various arguments could be made but it is likely that the samples it misclassifies are ambiguous and hold weak correlation, hinting at the serendipity of discovery rather than being purposely excavated in the field. Nevertheless, one way this can be resolved, through literature and our own approach, is by merging results of an additional model.

4.5 Spatial Discovery Model

Due to suboptimal Recall in our Transfer-learning approach, we introduced the supplementary model to enhance predictability. To produce an actionable map, we will first explore the model through Hold-Out before using K-Fold for global predictions.

Our dataset is taphonomic-focused, thus we should identify and prune any unproductive features to improve implicit Discovery-based learning.

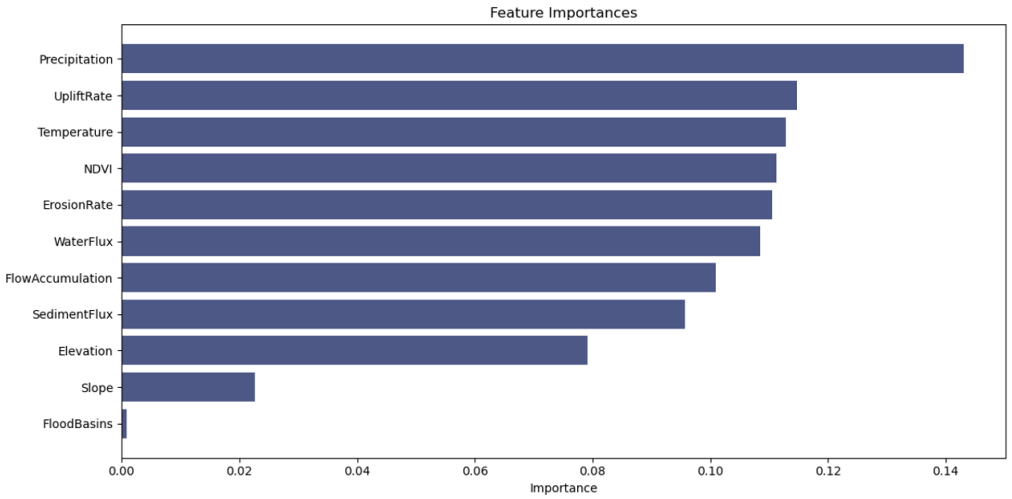

Figure 4.19: All Features, bar Flood-Basins/Slope, present Predictive Significance.

Feature importance shows that we can remove Flood-Basins and Slope as they are almost nominal. Almost identical to Figure 4.15. This could be drawn to spatial resolution and an issue that this study cannot investigate into due to its micro-fidelity improvements.

Anemone et al. (2011) remark that isolating cells with verticality (Slope) demonstrates better prospecting, however, they work on a fixed region at a higher spatial resolution. Our study looks at a much lower resolution, hence aggregated slope values and therefore reduction in predictability.

Features	Labels(s)
Precipitation	Full Occurrence Map (Figure 4.16)
Elevation	-
Temperature	-
Water Flux	-
Sediment Flux	-
Erosion Rate	-
Uplift Rate	-
Flow Accumulation	-
NDVI	-

Table 4.31: Pruned Experimental Dataset.

4.5.1 Creating Spatial Samples

We split the raster into 'cubes' of cells to create an array of spatial samples, as shown in Listing 4.4.

```
# Function to subdivide an even map to produce smaller trainable elements
def subdivideMap(featuresMap, labelMap, size, featuresCount):
    height, width = featuresMap.shape[:2]

    # Calculate number of cubes that can be made
    cubeCount = height // size * width // size

    # Create new arrays to hold the subdivided data
    subdivFeatures = np.zeros((cubeCount, size, size, featuresCount))
    subdivLabel = np.zeros((cubeCount, size, size))

    # Extract cubes
    idx = 0
    for i in range(0, height, size):
        for j in range(0, width, size):
            # Extract cube from features map
            subdivFeatures[idx] = featuresMap[i:i + size, j:j + size, :]

            # Extract corresponding label cube
            subdivLabel[idx] = labelMap[i:i + size, j:j + size]

            # Increment
            idx += 1
```

```
24
25      return subdivFeatures, subdivLabel
26
27  # Subdivide the features and labels into 10x10 resolution grids
28  subdivMapFeatures, subdivMapLabels = subdivideMap(mapFeatures, mapLabels,
        cubeSize=10, features)
```

Listing 4.4: Dividing Raster into Cubes Samples (Github).

We selected a dimension of 10×10 as it provides sufficient spatial resolution to assess spatial feasibility through multiple convolutional layers. It also balances detail and stability as smaller cubes add noise while larger cubes risk losing essential spatial features.

Additionally, we must choose a *cubeSize*(C_s) that evenly divides the raster dimensions $1441 - 1 \times 721 - 1$. This constraint can be expressed as follows:

$$\frac{1440}{C_s} \in \mathbb{Z}, \quad \frac{720}{C_s} \in \mathbb{Z}$$

4.5.2 CNN Model

Unregularised Model

Layer	Output Shape	Parameters
Input	(None, 10, 10, 9)	-
Conv2D	(None, 10, 10, 192)	15,744
Max Pooling	(None, 5, 5, 192)	0
Conv2D	(None, 5, 5, 384)	663,936
Max Pooling	(None, 2, 2, 384)	0
Conv2D	(None, 2, 2, 512)	1,769,984
Max Pooling	(None, 1, 1, 512)	0
Flatten	(None, 512)	0
Fully Connected 1 (ReLU)	(None, 512)	262,656
Fully Connected 2 (ReLU)	(None, 256)	131,328
Fully Connected 3 (ReLU)	(None, 128)	32,896
Fully Connected 4 (Sigmoid)	(None, 1)	129
	Total	**2,876,673**

Table 4.32: Unregularised CNN.

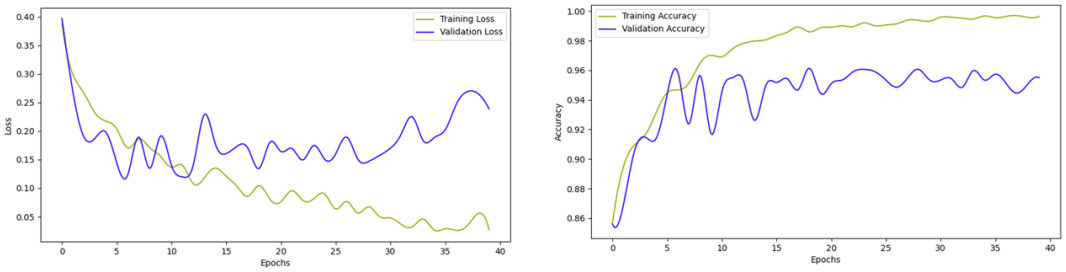

Figure 4.20: (Model 4.32) training instability observed in early model iterations, due to high learning rates and no regularisation.

The consistent spikes shown in Figure 4.20 suggest that the learning rate and/or batch size may need to be reduced to improve stability in training.

Test Set	**Binary Threshold**	**Metric**	**Unreg CNN**	**Random Guessing**
[8051 450]	0.30	Accuracy	0.9464	0.9014
	”	Precision	0.4933	0.0526
	”	Recall	0.3737	0.0505
	”	F1-Score	0.4253	0.0515
	N/A	Confidence	0.4895	0.0526
	0.30	AUC-ROC	0.8752	0.5000

Table 4.33: (Model 4.32) demonstrating Statistical Power despite overfitting.

Although classification accuracy is high, it is addressing the class imbalance. The inherent benefit in this model is to provide a large spatial extent to improve fossil prospecting, thus accuracy is not of importance.

Regularised Model

We then introduced regularisation with dropouts in each Convolution-Block, kernel regularisation and a significantly reduced learning rate:

Layer	Output Shape	Parameters
Input	(None, 10, 10, 9)	-
Conv2D	(None, 10, 10, 192)	15,744
Max Pooling	(None, 5, 5, 192)	0
Dropout	(None, 5, 5, 192)	0
Conv2D	(None, 5, 5, 384)	663,936
Max Pooling	(None, 2, 2, 384)	0
Dropout	(None, 2, 2, 384)	0
Conv2D	(None, 2, 2, 512)	1,769,984
Max Pooling	(None, 1, 1, 512)	0
Dropout	(None, 1, 1, 512)	0
Flatten	(None, 512)	0
Fully Connected 1 (ReLU)	(None, 512)	262,656
Dropout	(None, 512)	0
Fully Connected 2 (ReLU)	(None, 256)	131,328
Fully Connected 3 (ReLU)	(None, 128)	32,896
Fully Connected 4 (Sigmoid)	(None, 1)	129
	Total	**2,876,673**

Table 4.34: Regularised CNN.

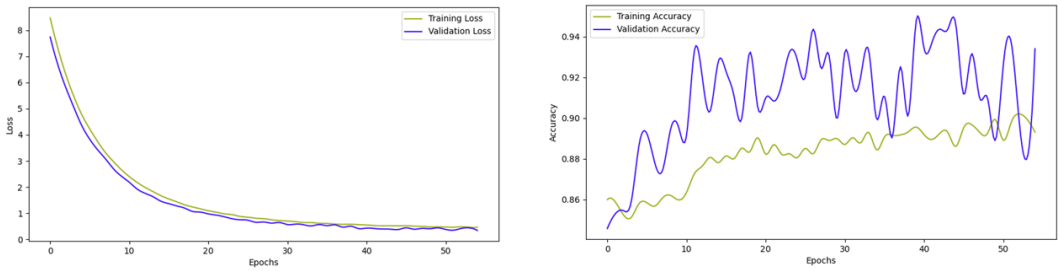

Figure 4.21: (Model 4.34) demonstrating stable loss but unstable accuracy.

Training exemplifies improvements and robustness in the loss curve. Yet despite gradual learning, accuracy remains unresolved.

Test Set	Binary Threshold	Metric	Reg CNN	Unreg CNN	Beaten?
[8051 450]	0.30	Accuracy	0.8870	0.9464	No
	”	Precision	0.2956	0.4933	No
	”	Recall	0.8182	0.3737	Yes
	”	F1-Score	0.4343	0.4253	Yes
	N/A	Confidence	0.3069	0.4895	No
	0.30	AUC-ROC	0.9400	0.8752	Yes

Table 4.35: Comparison between Regularised (Model 4.34) and Unregularised (Model 4.32) CNNs.

These results demonstrate a notable improvement in predictability, particularly in Recall. Lowering the threshold biases the model toward Recall, increasing false positives. If the model lacks strong predictive power, Precision remains low, making threshold adjustments less impactful.

Tuned Model

Once we tuned regularisation and learning rate we found correct training and subsequently, persuasive metrics.

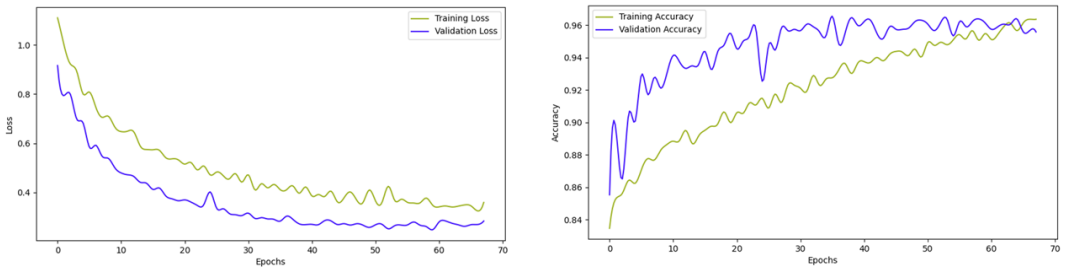

Figure 4.22: (Model 4.34) Tuned demonstrating appropriate training.

Test Set	Binary Threshold	Metric	Tuned CNN	Untuned CNN	Beaten?
[8051 450]	0.30	Accuracy	0.9239	0.8870	Yes
	”	Precision	0.3825	0.2956	Yes
	”	Recall	0.7071	0.8182	No
	”	F1-Score	0.4965	0.4343	Yes
	N/A	Confidence	0.3812	0.3069	Yes
	0.30	AUC-ROC	0.9224	0.9400	No
	”	IoU	0.3182	0.3274	No

Table 4.36: (Model 4.34) Tuned vs. Untuned Comparison.

There is a minor deficit to Recall, this is tolerable as the increase in Precision indicates enhancements in distinguishing positives from false positives.

4.5.3 Results: Maximising with Threshold-Moving

K-Fold Cross-Validation was subsequently used to train and predict the entire dataset. Following that, we tuned the threshold to maximise IoU.

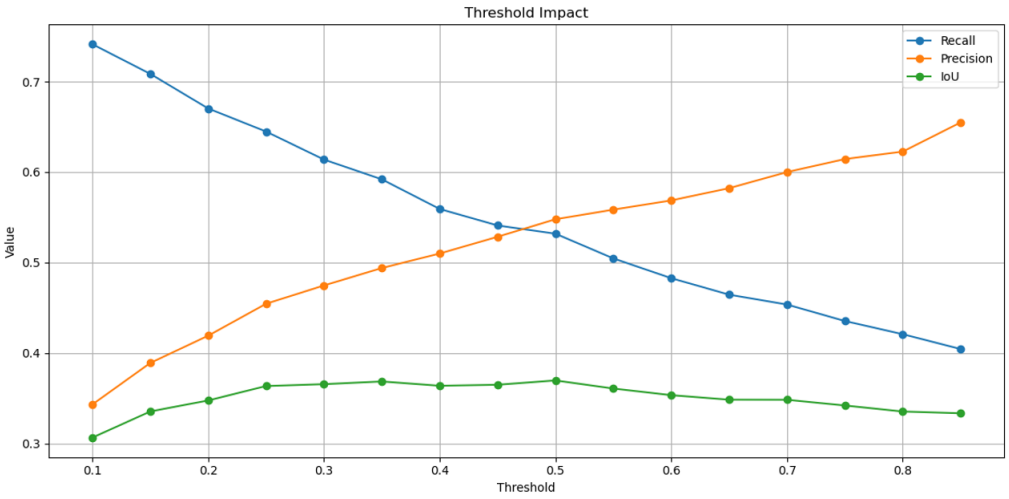

Figure 4.23: Threshold-Moving at 0.1 increments.

The model performs optimally when the threshold is set to 0.5. While this sacrifices some Recall, it maximises IoU and significantly improves Precision. K-Fold provides a more generalised performative outlook for smaller datasets and observe this productivity in our metrics in Table 4.37.

Test Set	Metric	K-Fold	Hold-Out	Beaten?
[8051 450]	Accuracy	0.9520	0.9239	Yes
	Precision	0.5478	0.3825	Yes
	Recall	0.5319	0.7071	No
	F1-Score	0.5397	0.4965	Yes
	Confidence	0.4499	0.3812	Yes
	AUC-ROC	0.9207	0.9224	Yes
	IoU	0.3696	0.3182	Yes

Table 4.37: (Model 4.34) K-Fold vs. Hold-Out Comparison.

With respect to improving performance, spatial aggregation reduced resolution to a point where further layers may cause ambiguity in feature extraction, thus enhancements would stem from higher-resolution data rather than introducing more model complexity.

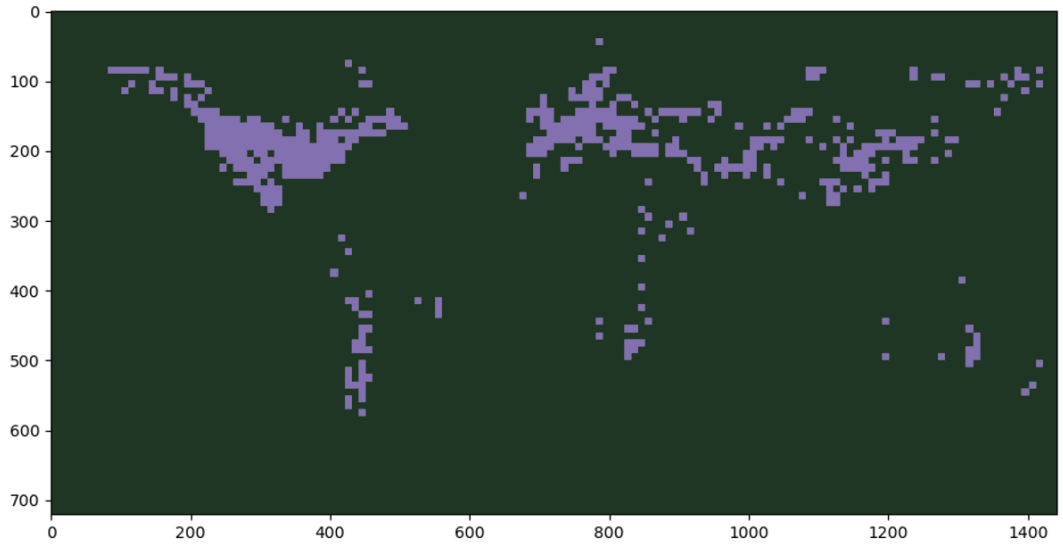

Figure 4.24: Cross-Validated Map (Purple = Fossiliferous Significance, Green = N/A).

To conclude, we saved the output predictions, as shown in Figure 4.24.

4.6 Harmonic Results

Our final results of the implementation are fabricated through superimposing the Transfer-Learned output (Figure 4.30) on our spatial output (Figure 4.24). We deduced the optimal weight by employing the same Threshold-Impact analysis (see Figure 4.23).

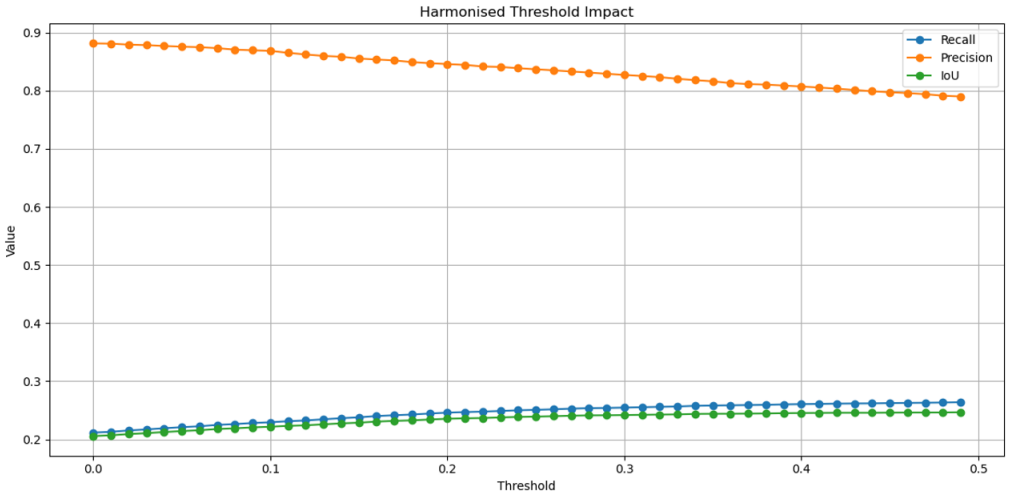

Figure 4.25: Threshold-Moving at 0.01 increments. Plateau at approximately 0.4.

Listing 4.5 shows the superimposition implementation.

```
# Apply the recall boosted values elementwise
adjustedPredictions = np.copy(globalPredictionsRaw)
# For every cell that has been boosted by the spatial model, increase the
    prediction by the harmonisation weight
adjustedPredictions[cellRecallBoostLabels == 1] += 0.4
# Ensure the adjusted predictions do not exceed the maximum of 1
adjustedPredictions = np.clip(adjustedPredictions, 0, 1)
```

Listing 4.5: Superimposing with Weight = 0.4 (Github.)

In drawing findings to close, we find that we have improved Recall, F1-Score, and IoU upwards of 5% across all classes through this harmonisation.

Binary Threshold	**Metric**	**Harmonised All Fossils**	**Harmonised Cretaceous**	**Harmonised Jurassic**	**Harmonised Triassic**
0.90	Accuracy	0.7328	0.8269	0.8524	0.8397
”	Precision	0.8073	0.7884	0.8233	0.8091
”	Recall	0.2606	0.4206	0.5214	0.4694
”	F1-Score	0.3940	0.5485	0.6385	0.5941
N/A	Confidence	0.5335	0.4976	0.5252	0.5196
0.90	AUC-ROC	0.8569	0.8978	0.9133	0.9135
”	IoU	0.2533	0.3779	0.4689	0.4226
	Dataset Ratio	2:1	3:1	3:1	3:1

Table 4.38: Harmonic Results.

To conclude implementation: we performed viability analysis, elected a preservation model, transfer-learned said model, and finally merged preservation and discovery predictions, producing our harmonic output, to ameliorate performance.

Chapter 5

Discussion and Evaluation

5.1 Hypothesis 1

The research found compelling evidence supporting deep-time temporal modelling as a viable predictive approach. It demonstrated convincing predictions for several geological eras (4.26), retained statistical power and considerably higher metrics when compared to non-temporal models (i.e, logistic regressions). Yet, when it came to the transfer-learned model underperformed - conceivably due to the proposed TP/TN sampling methodology as we dealt with a smaller dataset.

5.1.1 Is the preservation model sufficient?

To which at the time of this study, we concur with existing literature: no. Though the study investigated deeply at a broad angle and demonstrated empirical viability, but weak Recall (4.4) evince struggling in independency.

Could we assume linear-scalability in this issue? - to which our analysis implores future investigation. The focus of this study was to garner an analytical view, and having proving our hypotheses at a principle level, granular studies would indeed scale appropriately.

5.2 Hypothesis 2

The final predictive model was able to present mostly accurate predictions, it demonstrated strong generalisation capability (4.38). Spatial accuracy was found to be acceptable with

our best IoU of 0.4689 and worst of 0.2533 (4.38). However the temporal model was unable to resolve Recall to a level at which we could support its reliability (4.30).

5.2.1 Is the discovery model sufficient?

With reference to its purpose: yes. It functioned above statistical power and raised scores by $\approx 5\%$. Though at a larger scope, the isolated Recall of 0.5478 and Precision of 0.5319 (4.37) shows the inability to hold its own, it remains a viable approach with ample avenues for further exploration.

5.3 Research Q1: Will changing temporal steps impact the model's predictive power?

Analysis in 4.2.3 suggested that we would exhibit substantial improvements with reductions to time-steps, though narrowing the point of plateau could not be conducted due to dataset constraints. The nature of approximate data is to be synthetically accurate but not entirely ground true, hence a threshold at which a balance between generalisation and resolution would be necessary. Regardless, findings across the study (4.3.3, 4.4.2) prove these patterns extrapolate for RNNs.

5.4 Research Q2: Have we elected meaningful feature extractions and what was their influence on the models?

Our elected features continuously proved viability, most of which had consistent influence and importances (4.5, 4.15, 4.19). We found that some proposed features such as the 'Death Signal' feature demonstrated low predictive importance and was not retained. We pruned Slope and Flood-Basins from the CNN experiments to prevent training confusion as they could not substantiate justifiable importance scores.

5.4.1 Did we find any distinctions between Discovery and Preservation?

We found that the share a vast amount of overlap in how the same features can be used to induce analogous inferences. As they share an umbrella of factors that affect fossilisation,

it is an expected result, however to the degree of likeness we were not anticipating in this study.

5.5 Results Evaluation

With respect to our outcomes, our study's results lie between theoretical and complete empirical success.

Goal	Metric Threshold	Outcome
Ensure both models can beat baseline metrics	Greater than random guessing figures.	Met
Ensure both models achieve validation metrics	AUC-ROC $\geqq$ 0.85%. Precision / Recall $\geqq$ 0.95%.	Partially Met
Ensure the CNN model achieves persuasive spatial metrics	Spatial extent of $\geqq$ 0.30%.	Met
Harmonic Suitability Result	Identify $\geqq$ 0.90% of known fossil locations with high-confidence.	Not Met
Hypothesis 1 Evaluation	AUC-ROC $\geqq$ 0.80%. Precision / Recall $\geqq$ 0.90%. IoU $\geqq$ 0.20%.	Partially Met
Hypothesis 2 Evaluation	AUC-ROC $\approx$ 10-20% MoE. Precision / Recall $\approx$ 10-20% MoE. IoU $\geqq$ 0.50%.	Partially Met

Table 5.1: Outcome of Goals.

A consistent trend of high 80–95% AUC-ROC values confirms the overall effectiveness of our models, further complimented by beating all logistic models. The results highlight the efficacy of deep-time modelling to identify relevant signals for fossil locality prediction.

Unfortunately, our target Precision and Recall scores of 0.95 were not fully reached. When looking at predictive capability as a whole, we achieved results in the low 0.80s for Precision and substantially lower Recall (4.38). When investigating in forgiving conditions, we experienced similar Precision and upwards of 0.90s Recall (4.25). Regardless, these rates contextualise and validate real-world classification ability if somewhat below our desired goals.

Further confirmation is provided by spatial analysis. Despite a coarse prediction output, our CNN model obtained an IoU of 0.3697 (4.37), validating its capacity to generalise to broader regional patterns rather than simply overfitting.

Interestingly, success is observed where the model discovers regions of relevance that demonstrate strong predictive cohesion in both temporal and spatial aspects (4.30). Our results support pairing temporal with static models, as they lead to congenial integration, particularly in applications where data may be sparse or incomplete.

Despite the ambitiousness of our goals, and the inherent stochasticity of this task, we achieved a persuasive level of generalisation at our predictive resolution. In that sense, our main hypothesis can be accepted as verified to a confident degree. We would say that the results of this results could be used by field experts in prospecting new sites, thus implore inquiry into additional temporal paradigms at higher-resolutions to further justify real-world applicability.

5.6 Meaning within Broader Literature

Whilst prior studies use different methods/metrics, our findings show alignment with existing literature.

Our findings enhance the work of Block et al. (2016) by demonstrating that palaeographic macro-changes can validate the connections between environments and fossils. We achieve comparatively higher AUC-ROC scores, (> 0.63) for our preservation analysis, whilst also using unreliably dated occurrences. More-over the notion of combining taphonomic and discovery data to amplify predictions aligns with our unification findings. This strategy shows a promising direction for future prediction research.

Oheim (2007) highlighted the importance of preservation and discovery potentials. Our emphasis on temporal dynamics complements this perspective, demonstrating that time-series analysis can capture trends that static models will overlook. Ultimately, hinting at a synergetic relationship that future models can continue to improve.

Regarding neural network behaviour, Anemone et al. (2011) report scores of 84.21%, 98.85%, and 79.03% for Accuracy, Precision, and Recall respectively at a 0.95 threshold. Our results fall within a similar range, shown in Table 4.9. Though methodologies and experimental setups differ uniquely between the studies, the comparable performance highlights reliable model behaviour and methodologically sound investigations.

Chapter 6

Limitations and Future Work

6.1 The Pseudo-Absence Solution

We applied random sampling to garner a heuristic set of negatives, assuming they shared no geologic similarities, obtained from distinct environments. However, this approach risks sampling biases, preservative variability and temporal overprinting, thus potentially misrepresenting true absence.

Though it seems productive, we could not manifest consistently high metrics, thus we do not completely address the underpinning problem. Future studies can attempt validation at a localised scale to appropriate causation with sample depth. Otherwise, we can rule it out by gathering true-negative prospecting data - this would entail close access to field expert data.

6.2 Hyperparameter Search

We primarily wanted to deduce regularisation-focused values. The study investigated a broad array of models, thus it was unsustainable in the allocated time to exhaust further parameters/permutations in increased depth. Debatably, prolonged searching would not yield improvements for specific parameters such as batch size, optimisers and ADAM's epsilon - it should be noted that this was investigated but warranted no real room for tuning in this study.

6.3 Targets/Labels

Prediction targets endured some spatial aggregation, even at the highest resolution we explored. Methodologies for determining sedimentary regions were subject to spatial inaccuracies through dataset source and sampling bias through occurrence selection. Sampling evoked bias and whilst mitigations were such as clamping density and occurrences to a binary function were used, it did not absolve over/undersampling. An appropriate solution would be to quarry evenly distributed regions, in higher-resolutions, akin to Anemone et al. (2011).

Precise targets would be obtaining non-serendipitous occurrence data, ensuring validation is against accurately recorded data. With scalability proven, this could emphasise our hypotheses. Unfortunately due to time we could not validate the improvements this approach could make.

6.4 Dataset

Further features could be explored through extra physiographic derivations. Regarding the 'Death Signal' feature, there may be some inherent benefit in an ecological sense rather than preservation. All experiments highlighted temperature and precipitation as the highest importance features, future studies should refer to higher fidelity simulation data rather than relying on upscaling methods.

More-over we argue the most significant characteristic is the temporal interval. The balance between strong-features and interval-width could be disputed, a debate we cannot conclusively determine from this study.

6.5 Temporal Constraints

Temporal scope is subjective to the researcher in all literature and the appropriate length is that of which ensures all relevant information is incorporated into the experiments and analysis, for our study we focused on the classifications of the Mesozoic: Triassic, Jurassic, Cretaceous - to strike a balance between rich-sampling and data reliability. We expected aggregations in Pre-Mesozoic and Post-Mesozoic as a result.

An unavoidable constraint was the time-step interval, we were not limited by the historic

depth but rather the steps at which we sampled our features from. We would suggest further studies to probe smaller intervals as they ameliorated performance. See Appendix A.4.

6.6 Spatial Constraints

Although we explored consistency strategies (see Section 3.3.3), this issue stems from the resolution of the dataset. Recall that existing research also face data sparsity, which we similarly observe across our broader spatial extent. Adapting the working-space resolution to coarser degree involves extra computational effort, prolonged model training but ultimately has to produce higher yields (see Appendix A.3, A.4)

Chapter 7

Conclusion

In conclusion, we set out to investigate feasibility in fossil-presence prediction by leveraging deep-time data. We hoped to overcome intrinsic serendipity by refining prospecting beyond conventional methods. While previous work has explored machine-learning, much of it has focused on localised-cases or lacked a prominent temporal aspect, of which our study attempted to investigate.

First, we assessed effectiveness of deep-time modelling through benchmarking against baselines. Followed by predictions into simulated absence-free sedimentary regions, ensuring feature/target credibility. Then, we introduced geologic-focused LSTMs to construct environmental relationships. Transfer-learning was employed to preserve geological information while adjusting to true occurrences. Lastly, a spatial model was devised to improve results through a harmonic output. Evaluation confirmed the temporal principle but ultimately, fell short in a few quantified goals.

Despite observed strong performances, limitations remain. Sparsity in data, specifically occurrences, is an unavoidable uncertainty, and while we attempted to mitigate bias, we could not fully resolve imbalances. Yet, our findings show that preservation modelling is a viable avenue for improving fossil discovery. Future work could involve a wider search of features, richer palaeo-reconstructions, reliable occurrence reconnaissance and advanced models.

In the end, the study shines a light on the potential robustness of temporal modelling. Though targeted as a replacement, it can function as a robust tool for experts to reduce exploration costs, justify new sites and advance scientific understanding of lost biodiversity on our planet.

References

Anemone, R., Emerson, C. and Conroy, G. (2011), 'Finding fossils in new ways: An artificial neural network approach to predicting the location of productive fossil localities', *Evolutionary anthropology* **20**, 169–80.

Block, S., Saltré, F., Rodríguez-Rey, M., Fordham, D. A., Unkel, I. and Bradshaw, C. J. A. (2016), 'Where to dig for fossils: Combining climate-envelope, taphonomy and discovery models', *PLOS ONE* **11**(3), 1–16.

Bogdanovich, N., Kozlova, E. and Karamov, T. (2021), 'Lithological and geochemical heterogeneity of the organo-mineral matrix in carbonate-rich shales', *Geosciences* **11**(7).
URL: *https://www.mdpi.com/2076-3263/11/7/295*

Chorlton, L. B. (2007), 'Generalized geology of the world: Bedrock domains and major faults in gis format: A small-scale world geology map with an extended geological attribute database'.
URL: *https://ostrnrcan-dostrncan.canada.ca/entities/publication/9b769ab3-5525-4f3f-984d-493dafbcbf0d*

d'Oliveira Coelho, J., Anemone, R. and Carvalho, S. (2021), 'Unsupervised learning of satellite images enhances discovery of late miocene fossil sites in the urema rift, gorongosa, mozambique', *PeerJ* **9**.

GPlates (n.d.), 'Gplates/gplately: Gplately is a python package to interrogate tectonic plate reconstructions.'.
URL: *https://github.com/GPlates/gplately*

Hendricks, J. R. (2024), 'Geologic time scale'.
URL: *https://earthathome.org/geologic-time-scale/*

Hlusko, L. (2010), 'Fine-tuning paleoanthropological reconnaissance with high-resolution satellite imagery: The discovery of 28 new sites in tanzania'.

Hochreiter, S. and Schmidhuber, J. (1997), 'Long short-term memory', *Neural Comput.* **9**(8), 1735–1780.
URL: *https://doi.org/10.1162/neco.1997.9.8.1735*

Idrees, H. (2024), 'Understanding the differences in neural networks'.
URL: *https://medium.com/@hassaanidrees7/ann-vs-cnn-vs-rnn-vs-lstm-understanding-the-differences-in-neural-networks-94486cbb6d5a*

Joseph, M. and Elleithy, K. (2020), 'Digit recognition based on specialization, decomposition and holistic processing', *Machine Learning and Knowledge Extraction* **2**, 271–282.

Karim, M. R. and Menshawy, A. (2018), 'Deep learning by example'.
URL: *https://www.oreilly.com/library/view/deep-learning-by/9781788399906/9eb6f643-5cab-4b89-9e20-5b673754868a.xhtml*

Keras (n.d.), 'Keras documentation: The tuner classes in kerastuner'.
URL: *https://keras.io/keras$_t$uner/api/tuners/*

Malakhov, D., Dyke, G. and King, C. (2009), 'Remote sensing applied to paleontology: exploration of upper cretaceous sediments in kazakhstan for potential fossil sites', *Palaeontologia Electronica* **12**.

Matthews, K. J., Maloney, K. T., Zahirovic, S., Williams, S. E., Seton, M. and Müller, R. D. (2016), 'Global plate boundary evolution and kinematics since the late paleozoic'.
URL: *https://doi.org/10.5281/zenodo.10526157*

Myers, C. E., Stigall, A. L. and Lieberman, B. S. (2015), 'Paleoenm: applying ecological niche modeling to the fossil record', *Paleobiology* **41**(2), 226–244.

Novacek, P., Novacek, M. and of Natural History, A. M. (2001), *The Biodiversity Crisis: Losing what Counts*, An American Museum of Natural History book, New Press.
URL: *https://books.google.co.uk/books?id=MIV9QgAACAAJ*

Oheim, K. B. (2007), 'Fossil site prediction using geographic information systems (gis) and suitability analysis: The two medicine formation, mt, a test case', *Palaeogeography, Palaeoclimatology, Palaeoecology* **251**(3), 354–365.
URL: *https://www.sciencedirect.com/science/article/pii/S0031018207002362*

Phillips, S. J., Anderson, R. P. and Schapire, R. E. (2006), 'Maximum entropy modeling of species geographic distributions', *Ecological Modelling* **190**(3), 231–259.
URL: *https://www.sciencedirect.com/science/article/pii/S030438000500267X*

Salles, T., Husson, L., Lorcery, M. and Boggiani, B. H. (2022), 'Paleo-physiography project: Cuahsi hydroshare'.
URL: *http://www.hydroshare.org/resource/0106c156507c4861b4cfd404022f9580*

Scotese, C. R., Vérard, C., Burgener, L., Elling, R. P. and Kocsis, T. (2024*a*), 'Phanerozoic-scope supplementary material to "the cretaceous world: Plate tectonics, paleogeography, and paleoclimate" from the paleomap project'.
URL: *https://zenodo.org/records/10659112*

Scotese, C. R. and Wright, N. M. (2018), 'Paleomap paleodigital elevation models (paleodems) for the phanerozoic'.
URL: *https://doi.org/10.5281/zenodo.5460860*

Scotese, C., Vérard, C., Burgener, L., Elling, R. and Kocsis, (2024*b*), 'The cretaceous world: Plate tectonics, paleogeography, and paleoclimate', *Geological Society, London, Special Publications* **544**.

Seton, M., Müller, R., Zahirovic, S., Gaina, C., Torsvik, T., Shephard, G., Talsma, A., Gurnis, M., Turner, M., Maus, S. and Chandler, M. (2012), 'Global continental and ocean basin reconstructions since 200ma', *Earth-Science Reviews* **113**(3), 212–270.
URL: *https://www.sciencedirect.com/science/article/pii/S0012825212000311*

Stryker, C. (2024), 'What is a Recurrent Neural Network (RNN)? — IBM — ibm.com', https://www.ibm.com/think/topics/recurrent-neural-networks.

Torsvik, T. H., Van der Voo, R., Preeden, U., Mac Niocaill, C., Steinberger, B., Doubrovine, P. V., van Hinsbergen, D. J., Domeier, M., Gaina, C., Tohver, E., Meert, J. G., McCaus-

land, P. J. and Cocks, L. R. M. (2012), 'Phanerozoic polar wander, palaeogeography and dynamics', *Earth-Science Reviews* **114**(3), 325–368.
URL: *https://www.sciencedirect.com/science/article/pii/S0012825212000797*

Wang, Y. (2024), 'Research on the impact of fossil record on weather and climate prediction', *Theoretical and Natural Science* **70**, 38–43.

Wills, S., Choiniere, J. and Barrett, P. (2017), 'Predictive modelling of fossil-bearing locality distributions in the elliot formation (upper triassic–lower jurassic), south africa, using a combined multivariate and spatial statistical analyses of present-day environmental data', *Palaeogeography, Palaeoclimatology, Palaeoecology* **489**.

Wright, N., Zahirovic, S., Müller, D. and Seton, M. (2013), 'Towards community-driven paleogeographic reconstructions: integrating open-access paleogeographic and paleobiology data with plate tectonics'.

Ye, S. and Peters, S. E. (2023), 'Bedrock geological map predictions for phanerozoic fossil occurrences', *Paleobiology* **49**(3), 394–413.

Zhang, A., Lipton, Z., Li, M. and Smola, A. (2021), 'Dive into deep learning'.

Appendix A

Appendix

A.1 Source Code and Datasets

To meet our goal of redistribution, the cleaned and imputed datasets are available on the Github repository in NPY format.

- Code (Version Controlled via Github)
- Complementary Datasets (also available on Github)

The complete list of data sources used for the study.

- PaleoBioDB (Fossil Record) available here.
- Salles et al. (2022) Paleo-Physiography Project dataset.
- 2025 Landsat NDVI (Vegetation Indices) from NOAA.
- Scotese et al. (2024*a*) supplimentary material available here.
- Matthews et al. (2016) rotation model available here.
- Chorlton (2007) global generalised geology dataset.

A.2 Software Used

- QGIS Desktop 3.34.11 (with SAGA Next Gen tools) was the GIS software used in this study, alternatives such as ARCGIS would be an acceptable substitute.

- Python 3.10 with TensorFlow 2.10 were used. (Refer to documentation README.MD for compatibility).

- GPlates was used to visually inspect Rotation Models and the Phanerozoic DEMs before conducting the study.

A.3 Hardware Limitations

- The maximum spatial resolution explored in this entire implementation is $0.25 \times 0.25°$, although the system is not limited to this resolution, spatial resolution expands computation time exponentially, therefore, to strike the best balance between iteration and fidelity, $0.25 \times 0.25°$ was elected.

- No cloud or remote services were used for any part of the study, all research was conducted on a CUDA enabled GPU (2080 Super) with the following RAM and CPU configurations: 32GB, Ryzen 9 3900x. Training periods averaged hours and Hyperparameter tuning reached upwards of 2 days.

- No pre-existing machine-learing model weights or predictive-models produced from other sources were used for any of this research.

A.4 Scope Limitations

- The study explored a scope of 50 time-steps at the range (0-245 Ma), this gave sufficient range to incorporate the Mesozoic whilst also keeping a low enough range to not prolong pre-processing, model training and investigate analysis. Datasets offered data that range to almost 100 time-steps however even at an analytical standpoint, feature extraction took 30-60 mins per feature depending on any spatial alignment strategies, therefore it was not time-feasible to explore any wider scope for this study.

A.5 Leakage Limitations

- Although all models adhered to correct workflows for data splitting, data was indirectly leaked through the Pre-Processing resampling step.

- Theoretically this was amplified through the CNN model as it scaled the dataset before distributing cells into their respective spatial block samples however, we do not observe any unexpected behaviour or improved/worsened results due to this.

A.6 Clarification on Time Abbreviations

- Megaannum (Ma), Million year (Myr) are used interchangeably, although they refer to a unit of time, Ma refers to a period whilst Myr refers to a duration of time.

A.7 Hyperparameter Values

Parameter	**Value**
epoch	37
dropout_1	0.3
dropout_2	0.4
$\text{kernelRegularizer}_1$	0.01
learningRate(ADAM)	0.00013562501444661616
batchSize	256

Table A.1: Tuned Parameters for Model 4.9.

Parameter	**Value**
epoch	50
lstm_1	512
lstm_2	288
dropout_1	0.05
dropout_2	0.05
dropout_3	0.05
dense_1	64
learningRate(ADAM)	0.00049723
batchSize	256

Table A.2: Tuned Parameters for Model 4.18.

Parameter	**Value**
epoch	58
dropout3_1	0.35
dropout4_1	0.05
dropout3_2	0.3
dropout4_2	0.1
dropout_3	0.1
$\text{kernelRegularizer}_1$	2.9287e-05
dropout_4	0.05
$\text{kernelRegularizer}_2$	7.0696e-05
dropout_5	0.25
learningRate(ADAM)	0.00046695
batchSize	256

Table A.3: Tuned Parameters for Model 4.24.

Parameter	**Value**
epoch	60
$\text{kernelRegularizer}_1$	4.7902e-05
dropout_1	0.15
$\text{kernelRegularizer}_2$	0.00084717
dropout_2	0.1
$\text{kernelRegularizer}_3$	0.0047338
dropout_3	0.45
$\text{kernelRegularizer}_4$	0.00055775
dropout_4	0.45
$\text{kernelRegularizer}_5$	0.0002917
$\text{kernelRegularizer}_6$	0.0060515
learningRate(ADAM)	3.6914e-05
batchSize	256

Table A.4: Tuned Parameters for Model 4.34.

Parameter	Value
epoch	85
$\text{kernelRegularizer}_1$	0.005
dropout_1	0.1
learningRate(ADAM)	0.0001
batchSize	32

Table A.5: Tuned Parameters for Model 4.29.

A.8 Supporting Figures

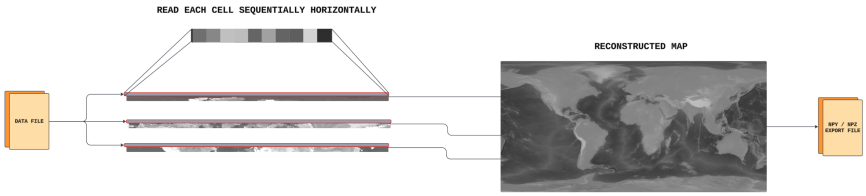

Figure A.1: CSV to NPY Conversion Pipeline.

A.9 Supporting Tables

Flag	Value
Lithology	NOT (Other, Metasedimentary, Metamorphic, Volcanic, Unknown)
Environment Type	Marine / Non-Marine
Taxon	All
Metadata	Coordinates, Environment Type, Geologic Time-bin

Table A.6: PBDB Query Parameters specified for Phanerozoic-scope occurrences.

Feature	Explanation
Precipitation	Mean annual precipitation rate for the interval.
Elevation	Estimated land elevation and ocean depth for the interval.
Temperature	Mean annual temperature for the interval.
Water Flux	100 Ka simulation for the rate of water movement.
Sediment Flux	100 Ka simulation for the rate of sediment movement.
Flood Basins (Lakes)	Flooded regions including lakes, water flow, outlets.
Erosion Rate	100 Ka simulation for the rate of erosion.
Uplift Rate	Rate of surficial adjustments derived through elevation deltas.
Slope	Verticality of a given area.
NDVI	Quantification of vegetative activity in a given area.
Death Signal	Occurrence pre-diagenesis activity for the interval.

Table A.7: Explanations of All Explored Features.

A.10 Changes to Harmonic Design

We initially proposed to improve scores using a stratigraphic cascade scoring system. However as we drew to the conclusion of using stratigraphy as a predictive target (see subsection 3.3.1), it became a redundant idea.

It should be noted that although this system has been proven in principle by existing work (e.g., (Myers et al., 2015)), attempting to use this scoring system with our experimental pipeline introduces biases that would inadvertently inflate performance. Furthermore the original intention was to improve performance if the models completely lacked spatial credibility, a circumstance we did not anticipate, and in which case the harmonic output satisfied.

A.11 Updates to Final Project

- All diagrams have been improved with en-largened text and higher-resolution than previous hand-in submissions.
- Chapter 1 has been revised to improve the understanding of motivations and aims with appropriate sources.

- Chapter 2 has been updated with some cuts to background research to improve cohesion in understanding in existing work and the gap this project is attempting to fill.
- Chapter 2 also includes literature in the fundamentals of RNNs as a supplementary explanation for why this method is suited for the proposed project.
- Chapter 3 features a simplified system proposal diagram.
- Chapter 3 talks in further detail about subtopics such as model choices to relate methodology to literature better.
- Chapter 3 also details the changes in the pipeline causing a shift in how the harmonic output is processed and why the previous design cannot integrate with the improved one.
- Chapter 3 also talks about methodological choices and further details in evaluation design as per feedback on previous submissions.
- Chapter 3 also details rewording of research questions and quantification goals to better fit the implemented approach.