

15 **Big Data, AI, and Deep Learning in Health Systems**

What Changes at Scale

Much of this book's material applies whether a dataset has hundreds or millions of records, but genuinely 'big' health data — nationwide HMIS transaction logs, electronic medical record systems across a hospital network, mobile phone call-detail records used as a proxy for population mobility, or large medical imaging archives — introduces distinct challenges characterised by the classic 'three Vs': volume (data too large for a single machine's memory, requiring distributed computing frameworks), velocity (data arriving continuously and requiring streaming rather than batch processing, as in real-time surveillance dashboards), and variety (heterogeneous formats — structured registers, free-text clinical notes, images — that must be integrated). Nigerian institutions are increasingly encountering all three, as DHIS2 instances mature, as hospital electronic health record adoption grows, and as mobile-health data collection scales.

Deep Learning

Neural Networks Beyond Linear and Logistic Regression

A neural network can be understood as a generalisation of the logistic regression of Chapter 8: instead of a single linear combination of predictors passed through a sigmoid function, a

neural network chains many such transformations across multiple layers, each capable of learning increasingly abstract, non-linear combinations of the input features, fitted by an algorithm (backpropagation with gradient descent) that adjusts all the network's weights to minimise a loss function. This additional flexibility comes at a cost: neural networks typically require substantially more data to train reliably than the classical models of Chapters 8-11, they are more prone to overfitting without careful regularisation, and they are considerably less directly interpretable — considerations that mean a neural network is often not the right tool for a modest-sized clinical cohort study, even though it may be the right tool for a large-scale imaging or text dataset.

Convolutional Neural Networks for Medical Imaging

Convolutional neural networks (CNNs), which learn spatial filters that detect edges, textures, and increasingly complex visual patterns through successive layers, are the dominant architecture for image-based diagnostic tasks — for example, classifying a blood smear image for malaria parasitaemia, grading a chest X-ray for tuberculosis-consistent findings, or, closer to the fingerprint-pattern (dermatoglyphic) research introduced in Chapter 12, potentially automating the classification of ridge patterns that are currently coded manually by trained researchers. A CNN deployed in a low-resource clinical setting must be evaluated not only on accuracy but on its performance across the range of image quality actually encountered in the field (variable smartphone camera quality, inconsistent lighting), since a model validated only on high-quality reference images can fail unpredictably outside that narrow condition.

Natural Language Processing for Clinical and Public Health Text

A large fraction of health information is recorded as unstructured free text — clinical notes, verbal autopsy narratives, social media

posts relevant to disease outbreak signal detection (an emerging area sometimes called digital epidemiology or infodemiology). Natural language processing (NLP) techniques ranging from simple keyword and rule-based extraction, through classical bag-of-words and TF-IDF representations feeding into the classifiers of Chapter 11, to modern transformer-based language models, allow this unstructured information to be converted into structured variables — for example, automatically extracting a probable cause of death category from a free-text verbal autopsy narrative, an application directly relevant to strengthening Nigeria’s vital registration system given the persistent under-registration and under-certification of deaths outside formal health facilities.

Practical Realism

When (Not) to Reach for Deep Learning

This chapter closes with a deliberately cautionary note that runs against a common assumption among newcomers to the field: for the great majority of structured, tabular health datasets of the kind most Nigerian researchers actually work with — a few hundred to a few thousand patients, a few dozen variables — a well-specified logistic regression, random forest, or gradient boosting model (Chapters 8 and 11) will typically match or outperform a deep neural network, train far faster, require far less data, and remain far more interpretable to a clinical or policy audience. Deep learning earns its complexity cost specifically on image, text, and very-large-scale structured data; choosing it by default for a modest tabular dataset is a common and avoidable error.

Worked Example

A Minimal CNN for Blood Smear Classification

A minimal CNN for classifying microscopic blood smear image patches as parasitised or uninfected (a simplified version of a real, published task) can be built with just a few convolutional and

pooling layers. On a labelled dataset of several thousand image patches split 80/20 into training and test sets, such a model can typically achieve a test accuracy in the region of 92-96%, though the precise figure depends heavily on image quality, staining consistency, and the diversity of the training set — a model trained on images from a single microscope and laboratory technician's staining technique will generally not generalise as well to images from a different facility, echoing the subgroup-performance caution raised in Section 11.5 and again in Section 15.3.

Figure 15.1 — Minimal CNN Architecture for Blood Smear Classification (Sec. 15.6)



Figure 15.1 — A minimal CNN architecture for blood smear classification (Section 15.6).

PYTHON example

```

# Python: a minimal CNN for blood smear patch classification with Keras
from tensorflow import keras
from tensorflow.keras import layers

model = keras.Sequential([
    layers.Rescaling(1./255, input_shape=(64, 64, 3)),
    layers.Conv2D(16, 3, activation="relu"), layers.MaxPooling2D(),
    layers.Conv2D(32, 3, activation="relu"), layers.MaxPooling2D(),
    layers.Flatten(),
    layers.Dense(32, activation="relu"),
    layers.Dropout(0.3),
    layers.Dense(1, activation="sigmoid") # 1 = parasitised, 0 = uninfected
])

model.compile(optimizer="adam", loss="binary_crossentropy", metrics=["accuracy"])
history = model.fit(train_images, train_labels, validation_split=0.2, epochs=15)
test_loss, test_acc = model.evaluate(test_images, test_labels)
print(f"Test accuracy: {test_acc:.1%}")

```

Worked Example

Extracting Structured Information from Verbal Autopsy Text

A simplified rule-based NLP pipeline for verbal autopsy narratives searches free-text descriptions for keyword patterns associated with specific probable causes of death — for example, flagging mentions of ‘fits’, ‘convulsion’, or ‘stiff neck’ together with ‘fever’ as suggestive of a possible meningitis-related death, or ‘yellow eyes’ combined with ‘dark urine’ as suggestive of a possible hepatic or haemolytic cause. Such rule-based approaches are far simpler than a full transformer-based clinical NLP model, but they are transparent, auditable by a physician reviewer, and often sufficient as a first-pass triage tool to prioritise which verbal autopsy narratives most need expert physician review — illustrating, in miniature, the same ‘match the tool to the task and the data volume’ principle argued for in Section 15.5.

PYTHON example

```
# Python: a minimal rule-based keyword flagging system for verbal autopsy text
import re

meningitis_keywords = ["convulsion", "fit", "stiff neck", "bulging fontanelle"]
hepatic_keywords = ["yellow eyes", "jaundice", "dark urine"]

def flag_probable_cause(narrative):
    text = narrative.lower()
    flags = []
    if any(k in text for k in meningitis_keywords) and "fever" in text:
        flags.append("possible_meningitis")
    if any(k in text for k in hepatic_keywords):
        flags.append("possible_hepatic")
    return flags if flags else ["unclassified_needs_review"]

sample_narrative = "The child had high fever for three days followed by convulsions and
stiffness of the neck."
print(flag_probable_cause(sample_narrative))
```

Transfer Learning

A Practical Shortcut for Small Image Datasets

Section 15.5 cautioned that deep learning typically requires large datasets to train reliably from scratch — a serious constraint for

a Nigerian research group with only a few hundred or thousand labelled clinical images. Transfer learning offers a practical middle path: starting from a large neural network already trained on a very large, general image dataset, and re-training only its final layers on a much smaller, task-specific dataset (such as a local blood smear image collection), leveraging the general-purpose visual features the network already learned (edges, textures, shapes) rather than learning them again from scratch. This approach can often achieve respectable performance on datasets an order of magnitude smaller than would be needed to train an equivalent network entirely from scratch, making deep learning meaningfully more accessible for resource-constrained Nigerian research settings.

Extended Case Study

An NLP Tool That Worked Well in English but Failed on Code-Switched Text

A team builds a keyword-based verbal autopsy triage tool, similar in spirit to Section 15.7's example, and validates it on a sample of narratives collected and transcribed in formal English, achieving good agreement with physician-assigned probable causes of death. When deployed more broadly, however, the tool's performance drops noticeably on narratives collected in communities where interviews were conducted in a mix of English and a local language (a linguistic pattern known as code-switching, extremely common in everyday spoken Nigerian English across many regions), because the transcribed narratives in these cases frequently use local-language terms for key symptoms (for example, a local term for 'convulsion' or 'jaundice') that do not appear anywhere in the tool's English-only keyword list, causing the rule-based system to systematically under-flag cases it would have correctly identified had the narrative used the equivalent English term.

This case study illustrates a specific and easily overlooked instance of the general subgroup-performance caution raised

throughout this book (Sections 11.5 and 16.3 in particular, now appearing in an NLP context): a tool validated on a linguistically non-representative sample of the population it will actually be deployed across can perform substantially worse for exactly the communities whose linguistic patterns differ most from the validation sample, which in this case likely correlates with rural, lower-literacy areas whose verbal autopsy data may be especially valuable precisely because facility-based cause-of-death information is least available there. The practical remedy — extending the keyword list to include locally common terms and phrases identified through direct consultation with the field interview teams who actually conduct these interviews in the affected communities, rather than relying solely on formal medical or English-language terminology compiled by the tool’s original developers — is straightforward in principle but requires the initial validation and deployment monitoring discipline to actually surface the problem in the first place.

The Cost of a False Positive Versus a False Negative in AI-Assisted Triage

Every classification threshold applied to a machine learning or deep learning model’s continuous predicted probability (Chapters 8, 11, and 15) implicitly encodes a judgment about the relative cost of a false positive versus a false negative, and this judgment should be made deliberately and transparently rather than left at a default (often 0.5) threshold chosen for no clinically meaningful reason. For a CNN-based blood smear screening tool (Section 15.6) intended to flag samples for confirmatory expert review, a false negative (a truly parasitised sample missed) may be considerably more costly than a false positive (an uninfected sample flagged for review, costing only a few minutes of an expert’s time), which argues for deliberately setting a lower probability threshold that favours sensitivity over specificity — a choice that should be made explicitly, in consultation with the clinical team

who will act on the tool's output, rather than left to whatever a generic 0.5 default happens to produce.

The Environmental and Resource Cost of Large AI Models

A consideration increasingly relevant to responsible AI use, and directly connected to this chapter's practical-realism argument in Section 15.5, is the computational resource cost of training and running large deep learning models: training a large neural network from scratch can require substantial computing hardware (often specialised graphics processing units, GPUs) and electricity that may be genuinely difficult or expensive to access reliably in many Nigerian institutional settings, and even running (not training) a large pre-trained model for routine inference can impose meaningful ongoing computational and cost burdens on an institution with limited infrastructure. This is a further, practical (rather than purely statistical) argument reinforcing Section 15.5's caution against reaching for deep learning by default: beyond the risk of overfitting on a modest tabular dataset, a large model's resource requirements may simply be a poor institutional fit compared to a classical statistical or machine learning model (Chapters 8 and 11) that can be trained and run reliably on far more modest, locally available computing infrastructure — an important practical consideration when choosing a tool intended for sustained, in-country institutional use rather than a one-off research demonstration.

Synthetic Data Generation for Privacy-Preserving Model Development

A promising and increasingly accessible technique for balancing the analytical value of a sensitive health dataset (Chapter 16) against its privacy risk is synthetic data generation: training a generative model on the real, sensitive dataset to learn its statistical structure, and then generating an entirely artificial

dataset of simulated ‘patients’ that preserves the real dataset’s statistical properties (variable distributions, correlations between variables) without corresponding to any single real individual. A well-constructed synthetic dataset can be shared far more freely for methods development, code-testing, and even training purposes than the original sensitive data, though it is not a complete substitute for the original when a specific finding must ultimately be confirmed on the real data, and it requires careful evaluation to ensure it does not inadvertently leak identifiable patterns from unusually distinctive real records.

Federated Learning

Training Models Without Centralising Sensitive Data

Chapter 16’s data sovereignty and privacy concerns are directly addressed by an emerging technical approach: federated learning, in which a predictive model is trained collaboratively across multiple institutions’ datasets (for example, several Nigerian teaching hospitals) without any institution’s raw patient data ever leaving its own servers — each institution trains the model locally on its own data, and only the resulting model updates (not the underlying data itself) are shared and aggregated centrally. This allows a model to benefit from a much larger, more diverse combined dataset than any single institution could provide alone, directly addressing the subgroup generalisation concerns raised throughout Chapters 11 and 16, while substantially reducing the data-sharing and sovereignty risks that would arise from physically centralising sensitive patient data from multiple institutions into one location.

Improving the CNN with Data Augmentation and Early Stopping

The minimal CNN in Section 15.6 trains for a fixed number of epochs on the raw images alone, which risks both overfitting to the specific training images and stopping either too early or too

late relative to when validation performance actually peaks. Two standard, low-cost improvements address this directly: data augmentation (randomly rotating, flipping, and adjusting the brightness of training images so the model sees many slightly varied versions of each blood smear, improving its robustness to the imaging variation discussed in Section 15.3) and early stopping (automatically halting training at the epoch where validation performance stops improving, rather than a fixed, arbitrarily-chosen epoch count).

PYTHON example

```
# Python: adding augmentation and early stopping to the Section 15.6 CNN
from tensorflow import keras
from tensorflow.keras import layers

augmentation = keras.Sequential([
    layers.RandomFlip("horizontal"),
    layers.RandomRotation(0.05),
    layers.RandomBrightness(0.1),
])

model = keras.Sequential([
    layers.Rescaling(1./255, input_shape=(64, 64, 3)),
    augmentation,
    layers.Conv2D(16, 3, activation="relu"), layers.MaxPooling2D(),
    layers.Conv2D(32, 3, activation="relu"), layers.MaxPooling2D(),
    layers.Flatten(), layers.Dense(32, activation="relu"), layers.Dropout(0.3),
    layers.Dense(1, activation="sigmoid"),
])
model.compile(optimizer="adam", loss="binary_crossentropy", metrics=["accuracy"])

early_stop = keras.callbacks.EarlyStopping(monitor="val_accuracy", patience=3,
                                             restore_best_weights=True)
model.fit(train_images, train_labels, validation_split=0.2, epochs=50,
          callbacks=[early_stop])
```

Chapter Summary

- Big health data introduces volume, velocity, and variety challenges beyond what classical statistical software handles comfortably.
- Neural networks generalise logistic regression via multiple non-linear layers, trading interpretability and data efficiency for flexibility.

- CNNs dominate image-based diagnostic tasks but must be validated under realistic field imaging conditions, not just curated reference images.
- NLP converts unstructured clinical/public health text (verbal autopsies, clinical notes) into structured, analysable variables, ranging from simple rule-based keyword flagging to modern transformer models.
- For typical modest-sized tabular health datasets, classical statistical and machine learning methods often outperform deep learning and should be the default choice.
- A simple, transparent rule-based NLP tool can be a sufficient and appropriately-scoped first-pass triage tool, rather than always requiring a state-of-the-art language model.

Mini-Project

Mini-Project 15: Choosing and Justifying the Right Tool for Scale

For a health data problem of your choice, make and justify an explicit decision about whether deep learning is warranted, and implement the appropriate simpler or more complex tool.

Step 1. Describe a health data problem involving either images, free text, or a large-scale structured dataset.

Step 2. Using the criteria of Sections 15.2 and 15.5, argue explicitly for or against using deep learning for this specific problem, citing data volume, format, and interpretability needs.

Step 3. If you argue against deep learning, implement and evaluate a classical method (from Chapters 8 or 11) instead. If you argue for it, implement a minimal version (following Section 15.6's structure).

Step 4. If your problem involves text, implement a simple rule-based keyword extraction (Section 15.7) as a baseline, even if you also try a more advanced method.

Step 5. Write a short paragraph on how you would evaluate your chosen tool's performance across different data sources or subgroups before trusting it operationally.

Deliverable: *A short justification for your tool choice, working code for your chosen approach, and a brief evaluation plan.*

End-of-Chapter Exercises

1. Give an example of each of the ‘three Vs’ (volume, velocity, variety) from a Nigerian health data context you are familiar with.
2. Explain why a CNN trained only on high-quality reference blood smear images might perform poorly when deployed using smartphone-captured images in a rural facility.
3. A colleague proposes using a deep neural network for a study of 300 patients and 12 clinical variables. What would you advise, and why?
4. Using the worked example in Section 15.6, explain why a CNN trained on images from a single laboratory might not generalise well to images from a different facility, even if test accuracy on the original dataset is high.
5. In Section 15.7, explain one advantage and one limitation of a rule-based keyword system compared to a more sophisticated NLP model, for the specific task of verbal autopsy triage.
6. Modify the Python code in Section 15.7 (in words or code) to add a third keyword category for a probable cause of death of your choosing, and describe what keywords you would include and why.
7. A donor wants to fund ‘an AI system’ for a Nigerian disease surveillance programme without specifying the task. Using this chapter’s arguments, write a short set of clarifying questions you would ask before recommending any specific tool.
8. Using Section 15.8, explain why transfer learning could make deep learning more feasible for a Nigerian research group with only 800 labelled clinical images.

9. What kind of general-purpose visual features does a pre-trained network already know, and why does re-using them save the need for a much larger local training dataset?
10. Would transfer learning fully solve the generalisation concern raised in Section 15.6 about a model trained on a single laboratory's images? Explain why or why not.
11. In the Section 15.9 case study, explain why the tool's drop in performance was specifically linked to code-switched narratives rather than a general decline in accuracy.
12. Why might this kind of error disproportionately affect exactly the communities whose verbal autopsy data is most valuable?
13. Describe the process you would use to identify local-language terms to add to the tool's keyword list, drawing on Section 15.9's proposed remedy.
14. Would you expect a transformer-based NLP model (Section 15.4) to be more robust to this specific problem than the rule-based tool? Explain your reasoning, including any new risks it might introduce.
15. Why is ongoing deployment monitoring, not just initial validation, necessary to catch a problem like the one in Section 15.9?
16. Using Section 15.10, explain why a lower classification threshold (favouring sensitivity) might be appropriate for a blood smear screening tool meant to flag samples for expert review.
17. Propose a clinical scenario where the opposite trade-off (favouring specificity over sensitivity) would instead be more appropriate, and explain why.
18. Why should the choice of classification threshold be made 'in consultation with the clinical team,' rather than left to a data scientist working alone?

19. Explain, in your own words, why a default threshold of 0.5 has no inherent clinical meaning.
20. If false negatives and false positives had genuinely equal costs in a given application, would adjusting the threshold away from 0.5 still be justified? Explain your reasoning.
21. Using Section 15.11, explain why training a large neural network from scratch could be a poor institutional fit for many Nigerian research settings, beyond the risk of overfitting discussed in Section 15.5.
22. Why does even RUNNING (not training) a large pre-trained model impose an ongoing resource burden, according to Section 15.11?
23. Propose one way a resource-constrained institution could still benefit from deep learning methods without needing to train a large model from scratch.
24. Explain the connection between Section 15.11's resource-cost argument and Section 15.8's discussion of transfer learning.
25. Why is 'a tool intended for sustained, in-country institutional use' an important framing when choosing between a classical model and a large deep learning model?
26. Summarise, in your own words, why this chapter argues that matching the tool to the task and data scale matters more than always choosing the most advanced available method.
27. Explain the common thread linking the code-switched text case study (Section 15.9) and the environmental cost of large AI models discussion (Section 15.11).
28. Why does this chapter insist on explicitly considering the relative cost of false positives versus false negatives (Section 15.10) before deploying an AI-assisted triage tool?

29. Propose an AI application from your own field where the cost of a false negative would clearly outweigh the cost of a false positive, or vice versa, and explain the practical implication.
30. If you had to advise a Nigerian institution considering a large, resource-intensive AI model, what questions from this chapter would you want them to answer first?
31. Using Section 15.13, explain how federated learning addresses the data sovereignty concerns raised in Chapter 16.
32. Describe what is shared between institutions in a federated learning setup, and, just as importantly, what is NOT shared.
33. Why might a model trained via federated learning across several institutions generalise better than a model trained on a single institution's data alone?
34. What practical or technical challenges might arise in coordinating a federated learning project across multiple Nigerian teaching hospitals?
35. How does federated learning relate to the subgroup generalisation concerns raised in Sections 11.5 and 11.7?