AIML-500 · WORKSHOP 2

Machine Learning vs. Deep Learning

Two real-world case studies, and why each approach fits — and the other does not

Sai Venkata Siddartha Darisi

The Question Behind the Question

The obvious way to answer “why is deep learning wrong for this problem?” is to say it is too complex, too data-hungry, or too expensive. That answer is usually true and it is almost never the real reason.

The real distinction is narrower and more useful: it is whether the features that predict the outcome can be written down by a human in advance.

> **THE GOVERNING PRINCIPLE** Classical machine learning is the right choice when a domain expert can enumerate the predictive features. Deep learning is the right choice when they cannot — when the signal lives in raw, high-dimensional data in a form no human can articulate as a list of columns. Everything else, including data volume and compute cost, follows from that.

This report examines two cases. In the first, the features are knowable, and a neural network has been tried on the same data and did not win. In the second, the features are not knowable, and the classical approach was tried for two decades before being abandoned. Neither “unsuitable” approach is unsuitable because it is theoretically incapable. Both are unsuitable because they were actually attempted and the evidence went the other way.

EXAMPLE 1 · MACHINE LEARNING

Customer Churn Prediction

1.1 The Problem and the Real-World Application

A telecommunications company wants to know which subscribers are likely to cancel their contract in the next month, so that retention offers can be targeted at the customers who will actually leave rather than sprayed across the whole base.

Real-world application. Churn prediction is one of the most widely deployed classical ML applications in industry. Published telecom deployments use logistic regression, decision trees, support vector machines, and ensemble methods against subscriber records. Documented results include a segmentation-plus-logistic-regression model reaching roughly 94% accuracy on telecom data (Tianyuan et al., cited in Machine Learning and Knowledge Extraction, 2025), and a prepaid-subscriber study in which logistic regression achieved the lowest misclassification rate (2.2%) of the algorithms tested. The most useful single source for this report is Frohböse (2020), who built a complete pipeline on the public IBM/Kaggle Telco dataset (~7,000 customers) and benchmarked KNN, logistic regression, random forest, and SVM — plus, deliberately as an add-on, a feed-forward neural network.

1.2 Why Classical ML Is Suitable

The features are already written down

This is the whole argument, and everything else is downstream of it. The data arrives as a table where every column is a meaningful, human-legible business fact:

- Contract type — month-to-month, one year, two year
- Tenure — how many months the customer has been with the company
- Payment method — electronic check, mailed check, bank transfer, credit card
- Monthly charges, total charges
- Services subscribed — internet, phone, streaming, tech support
- Demographics — dependents, partner, senior status

A telecom analyst did not need a neural network to suspect that month-to-month contracts churn more than two-year contracts. The domain knowledge came first; the model's job is to weight it, not to discover it. This is precisely the situation classical ML was designed for.

Interpretability is a requirement, not a nice-to-have

The model's output is not the deliverable. The deliverable is a retention campaign, and a retention team cannot act on “this customer scores 0.83.” They need to know why, because the intervention depends on the reason: a customer churning over price gets a discount, and a customer churning over service quality does not.

Logistic regression provides this natively. Its coefficients are the explanation. Frohböse's analysis of the Telco dataset states its drivers directly from the model's feature weights: high tenure is the strongest factor for not churning and the strongest feature overall; a month-to-month contract is the second biggest driver of churn;

and customers paying by electronic check churn at much higher rates than those on automatic payment methods. These are not post-hoc interpretations bolted onto an opaque model — they are the model.

The data volume does not support anything deeper

The canonical Telco churn dataset contains around 7,000 customer records with roughly twenty features. Even a large carrier's internal dataset is measured in millions of rows and dozens of columns — which sounds enormous but is tiny relative to what a deep network needs to learn representations from scratch. A model with more parameters than the dataset has meaningful structure will memorise noise, not signal.

1.3 Why Deep Learning Is NOT Suitable

> **THE STRONGEST EVIDENCE** This is not a theoretical objection. Deep learning has been tried on exactly this data, against exactly these baselines. Frohböse (2020) added a feed-forward neural network to his benchmark explicitly as an experiment, noting that the dataset is relatively small and neural networks generally require lots of training data. The network did score best — 0.7996 accuracy and 0.5948 F1 on the test set — but only marginally ahead of the classical models, and the author's own summary notes that no model's F1 rose much above 50%. The deep model paid a large cost in interpretability and training complexity for a gain measured in fractions of a point.

Honesty requires including the author's own caveat: his outlook section argues that real telecoms hold far more data than the ~7,000 records in this exercise, and that with such volumes neural networks could be properly trained to detect more complex patterns and achieve higher accuracies. That is fair — and it does not rescue the deep approach here, for two reasons. First, the comparison that matters is the one that was actually run, at the scale most churn projects actually operate at. Second, and more fundamentally, the interpretability requirement does not shrink as the data grows: a retention team still cannot act on an unexplained risk score, whether it came from seven thousand rows or seventy million.

There are no representations left to learn

Deep learning's central advantage is automatic feature extraction — discovering hierarchical representations in raw data that no human specified. But there is no hierarchy to discover in a table of contract types and monthly charges. The features are not raw. They are already the abstraction. A convolutional layer has nothing to convolve over; there is no spatial or sequential structure in a row of a customer database.

Asking a deep network to find latent features in tabular business data is asking it to do the one thing it is good at in the one domain where that thing has already been done.

The interpretability loss is disqualifying

Even if a neural network had achieved a meaningfully higher score, the result would be difficult to deploy. The retention team would receive risk scores with no explanation of the drivers. Worse, churn models are frequently used to justify spending on customer incentives — which means the model's reasoning is scrutinised by people who control budgets and who are entitled to ask why.

An accuracy improvement that cannot be explained to the business is often not an improvement at all. Practitioners working on this problem consistently report the same conclusion: interpretability outweighs complexity.

The costs are real and the benefit is not

A deep model on this problem introduces GPU training requirements, hyperparameter tuning, longer iteration cycles, and a harder production deployment — in exchange for a fraction of a point of F1 that may not survive the next quarter's data drift. That is not a trade an engineering team should make.

EXAMPLE 2 · DEEP LEARNING

Image Recognition — Diabetic Retinopathy Screening

2.1 The Problem and the Real-World Application

Diabetic retinopathy is a leading global cause of blindness and is preventable if caught early. Screening requires an ophthalmologist to examine a photograph of the retina for lesions — microaneurysms, haemorrhages, exudates, neovascularisation — and grade the severity. The bottleneck is specialists, and there are nowhere near enough of them.

Real-world application. IDx-DR received FDA approval in 2018, becoming the first autonomous AI diagnostic system cleared to detect diabetic retinopathy from retinal photographs without a physician interpreting the image. It applies a suite of CNN-based detectors trained to identify both normal anatomy — optic disc, fovea — and the pathological signs of the disease. EyeArt, from Eyenuk, received FDA clearance in 2020. Google Health deployed a deep-learning screening system across eleven clinics in Thailand, where roughly 4.5 million diabetic patients are served by about 200 retinal specialists; the system reports over 90% accuracy, described by the team as human-specialist level, returning a result in under ten minutes against a referral process that could otherwise take up to ten weeks.

2.2 Why Deep Learning Is Suitable

Nobody can write down the features

This is the mirror image of the churn problem, and it is the entire reason the approach differs.

Ask an ophthalmologist how they recognise a microaneurysm and they will say it looks like a small red dot. Ask them to define “small red dot” as a function of pixel values that distinguishes it from a haemorrhage, from a dust speck on the lens, from a normal vessel crossing, from an artefact of the camera flash, across every skin tone, every degree of cataract clouding, and every lighting condition in every clinic in the world — and they cannot. Not because they lack expertise, but because the knowledge is perceptual rather than propositional. They know it when they see it.

> **THE CORE DISTINCTION** In churn prediction, the domain expert's knowledge arrives as a list of columns. In retinal screening, the domain expert's knowledge arrives as twenty years of looking at retinas. The first can be handed to logistic regression. The second can only be handed to a model that learns its own representations from the raw pixels.

The CNN builds the feature hierarchy itself

A convolutional network learns in layers that correspond to increasing abstraction. Early layers detect edges and colour gradients. Middle layers assemble those into textures and shapes — vessel structures, lesion boundaries. Deep layers compose those into the clinically meaningful concepts: this is an exudate, that is neovascularisation, the overall grade is referable.

No human specified any of that. The hierarchy emerged from the data. That is not a convenience; it is the only known way to solve this problem at accuracy comparable to a specialist.

The input is raw and high-dimensional in exactly the way CNNs exploit

A retinal photograph is a grid of millions of pixels with strong spatial structure — nearby pixels are related, and the same lesion can appear anywhere in the frame. Convolutional architectures are built precisely around these two facts, through local receptive fields and weight sharing. The architecture matches the shape of the data.

Labelled data exists at the required scale

Deep learning's data requirement is met here in a way it is not met in churn prediction. Retinal screening programmes have accumulated enormous labelled corpora — one real-world validation study evaluated 311,604 retinal images from 23,724 patients across two Veterans Affairs health systems. Hundreds of thousands of expert-graded images is exactly the regime in which deep networks outperform everything else.

2.3 Why Classical ML Is NOT Suitable

It was tried for decades, and it lost

Before deep learning, automated retinal screening was attempted with hand-engineered feature extraction: researchers wrote explicit algorithms to detect microaneurysms by circularity and intensity thresholds, to segment vessels by morphological operations, to measure exudate area, and then fed those extracted numbers into classical classifiers such as SVMs.

These systems worked, in the sense that they produced results on the datasets they were tuned against. They did not achieve specialist-level accuracy, and they generalised badly. The reason is structural and it is worth stating precisely.

> **WHY HAND-ENGINEERED FEATURES FAIL HERE** Every hand-written feature detector is a human's guess about what matters, frozen into code. It captures what the engineer thought to look for and nothing else. A microaneurysm detector tuned on one population's images, one camera model, and one lighting setup encodes assumptions that break the moment any of those change. The classical approach cannot recover from an incomplete feature list — and the feature list is always incomplete, because the person writing it cannot articulate what they know.

Feeding raw pixels to a classical model does not work either

The alternative — skipping feature engineering and handing raw pixels directly to an SVM or a decision tree — fails for a different reason. A modest 1000×1000 retinal image is a million-dimensional input. Classical algorithms have no mechanism for exploiting the spatial relationship between adjacent pixels; to an SVM, pixel (500, 500) and pixel (500, 501) are simply two unrelated features among a million. Every model must relearn the concept of a lesion independently at every possible location in the frame, from scratch, with no shared structure.

This is the curse of dimensionality in its most literal form. The classical model has no inductive bias appropriate to images, and no amount of data compensates for that.

The stakes make “good enough” insufficient

This is a screening system whose false negatives are patients who go blind. The accuracy gap between hand-engineered classical pipelines and CNN-based systems is not a rounding error to be traded away for interpretability — it is the difference between a system regulators will approve for autonomous use and one they will not. Both FDA-cleared systems in this space are CNN-based. That is not a coincidence.

An honest qualification

Deep learning being the right approach does not mean the deployed systems are unproblematic. A multicentre head-to-head validation of seven automated screening algorithms — 311,604 retinal images from 23,724 veterans across two VA health systems — found that most performed similarly to or worse than the VA's human teleretinal graders in real-world conditions, and that the two algorithms achieving superior sensitivity did so at the cost of worse specificity (Diabetes Care, 2021). MIT Technology Review's account of Google's Thailand deployment reached a similar conclusion from the field: laboratory accuracy assessments only go so far, and the clinics' real conditions produced friction the accuracy numbers never predicted.

The lesson is not that deep learning is the wrong choice. It is that being the correct architecture for a problem does not make a system correct for a deployment. Lab accuracy and clinical usefulness are different properties, and the gap between them is where these systems actually fail.

Synthesis

	Churn Prediction	Retinal Screening
Approach	Classical ML — logistic regression, SVM, ensembles	Deep learning — convolutional neural networks
Input	~20 structured columns per customer	Millions of raw pixels per image
Who defines features	The domain expert, in advance	The model, from the data
Why the features are knowable	They are business facts — contract type, tenure, payment method	They are not. Perceptual expertise cannot be written as a rule list
Data scale	Thousands to millions of rows	Hundreds of thousands of expert-labelled images
Interpretability	Required — drives the retention intervention	Desirable but not decisive — accuracy governs approval
The other approach	Tried. NN scored 0.7996 accuracy / 0.5948 F1 — marginal gain, large cost	Tried for decades. Hand-engineered features never reached specialist accuracy

The rule that follows

Both examples resolve to the same question, asked in opposite directions.

If a domain expert can hand you the list of features, use classical machine learning. You already have the abstraction. A deep network would spend enormous capacity rediscovering what you were given for free, and would sacrifice the interpretability the business needs in order to do it.

If the domain expert can recognise the answer but cannot tell you how, use deep learning. That inability is the diagnostic signal. It means the predictive structure lives in the raw data in a form no feature list can capture — and learning representations from raw data is the one thing deep networks do that nothing else does.

The usual framings — that deep learning needs more data, or more compute, or is harder to interpret — are all true. But they are consequences, not causes. They describe the price of learning your own features. Whether that price is worth paying depends entirely on whether the features were available to you in the first place.

References

- Frohböse, F. (2020, November 24). Machine learning case study: Telco customer churn prediction. Towards Data Science. https://towardsdatascience.com/machine-learning-case-study-telco-customer-churn-prediction-bc4be03c9e1d/
- Alotaibi, M. Z., & Haq, M. A. (2024). Customer churn prediction for telecommunication companies using machine learning and ensemble methods. Engineering, Technology & Applied Science Research, 14(3), 14572–14578.

- Customer Churn Prediction: A Systematic Review of Recent Advances, Trends, and Challenges in Machine Learning and Deep Learning. (2025). Machine Learning and Knowledge Extraction, 7(3), 105. https://www.mdpi.com/2504-4990/7/3/105
- Heaven, W. D. (2020, April 27). Google's medical AI was super accurate in a lab. Real life was a different story. MIT Technology Review. https://www.technologyreview.com/2020/04/27/1000658/
- Multicenter, head-to-head, real-world validation study of seven automated artificial intelligence diabetic retinopathy screening systems. (2021). Diabetes Care, 44(5), 1168–1175. https://diabetesjournals.org/care/article/44/5/1168/138752/
- Artificial intelligence applications in diabetic retinopathy: What we have now and what to expect in the future. PMC. https://pmc.ncbi.nlm.nih.gov/articles/PMC11220221/ (IDx-DR FDA approval, 2018)
- The application of artificial intelligence in diabetic retinopathy: Progress and prospects. PMC. https://pmc.ncbi.nlm.nih.gov/articles/PMC11543434/ (CNN-based detector suite; EyeArt FDA clearance, 2020)
- Customer churn prediction in telecom sector using machine learning techniques. (2023). ScienceDirect. https://www.sciencedirect.com/science/article/pii/S2666720723001443

AI use disclosure: Claude (Anthropic) was used to structure this report and draft its prose. Source identification, the selection of examples, the governing argument, and all analytical judgements are my own.