Thesis Machine Learning
Deep Learning Risk Score Methodology

Table of Contents

1. Executive Overview

This document sets out the methodology used by Thesis Machine Learning to design, evaluate, deploy, and maintain classification models for covered assets and thematic sleeves. The methodology is rules-based, walk-forward, and designed to produce a consistent daily output, the **Deep Learning Risk Score, or DLR-Score**. Its purpose is to convert model outputs into a decision input that is auditable, governed, and implementable within an institutional investment process.

The methodology is presented as an end-to-end governed process rather than as an isolated model description. Data sourcing, point-in-time controls, target definition, feature selection, model evaluation, and production scoring are treated as parts of a single workflow. This is intended to reduce the risk that research results overstate what could reasonably be achieved in live use.

2. Plain English Summary

At a high level, our process is designed to answer a simple question: **based only on the information that would have been available at the time, how would our model evaluate an asset's risk profile?**

To do that, we maintain strict point-in-time integrity throughout the process. In practice, this means:

1. We use data inputs that reflect only the information that would have been available at the time.
2. We train and test models using a structure that allows performance to be evaluated without using information that models have not seen before.
3. Model selection is systematically assigned by top testing period model, and is only utilized after the testing period.
4. Training, testing, deployment, and replacement follow fixed, systematic schedules.
5. Outputs are produced and calculated through the same standardized process across all asset classes.

This structure is designed to ensure that both backtested and live results reflect the information that actually existed at the time. The result is a systematic, repeatable, and auditable output for clients.

3. Scope and Purpose

This methodology covers the model lifecycle from exposure selection through daily score generation. It sets out the standardized process used to produce our Deep Learning Risk Score ("DLR-Score") for each covered asset or thematic sleeve.

This document does not prescribe client-specific index construction overlays. Index rules, including investable universe, weighting, turnover controls, concentration limits, and mandate-specific constraints, are applied downstream per client requests, in a separate methodology document is provided for each index with the accompanying index rules. The DLR methodology is not impacted by index-specific construction requests.

The purpose of this document is to provide an externally reviewable reference for institutional partners, including index providers, ETF issuers, benchmark committees, diligence teams, and governance reviewers.

4. Covered Universe and Exposure Selection

Model development begins with the selection of an asset within the investable universe. This may include individual securities, ETFs, indices, amongst other exposures, provided the required history, data continuity, and data quality are sufficient to support disciplined model development.

Each covered exposure is maintained within its own governed model record. This allows lineage, validation history, eligibility, and replacement decisions to be assessed within the context of that exposure rather than across a pooled process.

5. Feature Design and Input Framework

Feature design is specific to the economic characteristics of each covered exposure. Inputs are selected because they have a plausible explanatory relationship to the behavior of the exposure under review. Depending on the asset type, inputs may include:

- price and trend measures
- volatility and dispersion measures
- valuation and cash flow metrics
- macroeconomic and cross-asset signals
- ETF holdings information (where relevant)
- other exposure-specific variables with a defensible economic rationale

The input set is not allowed to drift informally over time. For each training period, the methodology preserves the exact feature definition, scaling approach, and selected inputs used for model development and later scoring. This supports consistency between research, validation, and live use.

6. Historical Data and Point-in-Time Integrity

Historical coverage is generally targeted from 2008 onward where that history is available and economically meaningful. Shorter histories may be used where required by launch/IPO dates, data constraints, or structurally newer asset classes (such as digital assets).

Point-in-time integrity is the core principle of our methodology. Each observation is constructed using only information that would reasonably have been available at the time. In simple terms, we only train our models with data that would have been known on the date that is being analyzed.[1]

Monitoring records are retained to support audit review of date-source usage. These controls strengthen explicit documentation of a framework that was already designed to avoid forward-looking joins.

[1] For example, when integrating financial statement data, we anchor each report to its filing date rather than the period it covers. Even if we know today that a company announced earnings on a given date, our training data assumes we would not have access to the detailed results until they were formally filed. This ensures models are trained only on information that would have been available at the time, avoiding any reliance on knowledge learned after the fact. We apply this point-in-time discipline across all datasets to preserve the integrity of our results.

7. Target Definition and Classification Framework

During model development, each historical observation is labeled using the realized price move over a predefined horizon (from 't' to 't+h'). This ex post labeling step is used only to construct supervised training targets and is not available at inference time. In live inference, the model uses only information observable as of the decision date 't' to estimate the probability that an exposure is undervalued, fair-valued, or overvalued.

Threshold setting is performed through a structured search across candidate horizons and percentage-move levels. The selected combination is intended to maintain sufficiently balanced class distributions for stable training while remaining representative of the exposure's historical move profile over a certain number of days. In practical terms, this is a volatility-aware threshold calibration process tailored to each asset. It helps define how the model distinguishes between undervalued, fair-valued, and overvalued states for that asset. Once established, the core threshold parameters are carried forward unless a formal governance decision approves a change.

8. Model Development Framework

We utilize our proprietary deep learning classification framework to estimate the three class probabilities corresponding to undervalued, fair-valued, and overvalued states, on each NYSE trading day.

Development is staged as follows:

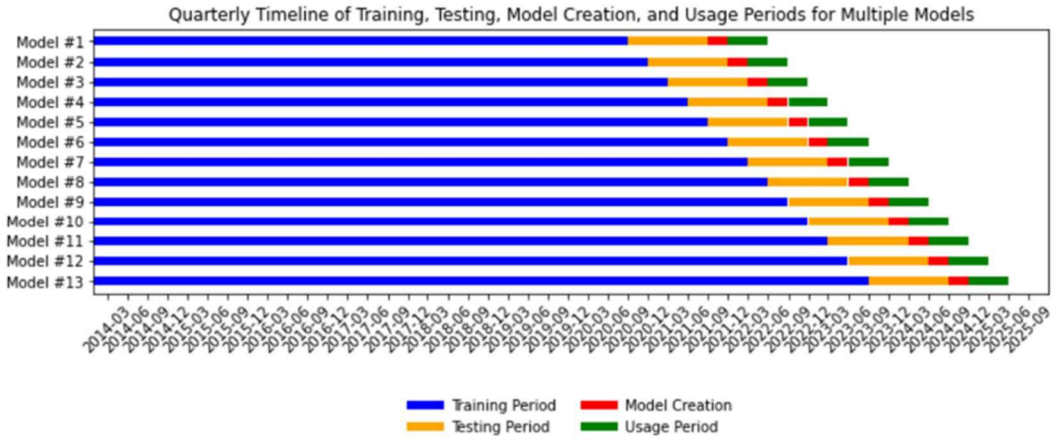
- Base model training on historical data available before each test window.
- For each training stage, the highest-ranked base model by F1[2] score is selected for the subsequent transfer-learning stage.
- Retention of model weights, architecture settings, selected features, scaling records, and performance records needed for reproducibility.

For each model vintage, the methodology preserves the relevant model files, architecture settings, input specifications, scaling parameters, and validation records needed to reconstruct lineage and reproduce results. Training and adaptation follow rule-based procedures rather than discretionary reconfiguration from period to period.

[2] F1 score is the harmonic mean of precision and recall, calculated as 2 × (Precision × Recall) / (Precision + Recall). In this methodology, F1 is measured out of sample on each model's forward testing window and weighted across the three classes to account for class imbalance. It serves as a governance measure of classification quality rather than a direct return forecast. Models must achieve an F1 score of at least 0.50 to be eligible for production, which reflects a minimum threshold at which the model demonstrates a consistent statistical edge in classification performance over random assignment during the testing period. Among active transfer-learning candidates, the two highest-F1 models are selected for live scoring.

9. Training, Testing, and Usage Windows

The methodology follows a walk-forward design on a rolling quarterly schedule. For each model vintage, the training dataset ends before the relevant test window begins. The test window is reserved for out-of-sample evaluation. The subsequent usage window is the period in which the score would have been available for live decision-making. This structure is designed to preserve strict separation between research, evaluation, and forward use, and to reduce information leakage, selection bias, and regime-specific overfitting.



Illustrative training, testing, model creation, and usage windows for successive model vintages.

A minimum 15 calendar-day freeze separates the end of testing from the beginning of usage. This freeze provides time for validation review, reproducibility checks, and production readiness before a model output can influence a live index or portfolio process.

Because models are introduced on a quarterly schedule and may remain eligible for six-month usage periods, four transfer-learning candidates may be eligible at the same time. The methodology therefore always evaluates eligible candidates within a governed schedule rather than assuming a single active model.

Plain English Emphasis

Our training, testing and usage window process is one of the most important controls in our methodology to ensure a fully systematic backtest and live utilization.

We separate the process into three distinct periods: training, testing, and usage. In the training period, the model learns from older historical data. In the testing period, the model is evaluated on a later period it has not seen before. Only in the usage period can the model finally be used to produce scores implemented in backtested and live results.

That separation matters because it helps ensure that the backtest reflects what could have happened in real time:

- It prevents data leakage[3] by ensuring information from the testing period is never used to train the model, which would unfairly inflate test performance.
- A model is not selected because it would have looked good in a later forward usage period. It is selected only because it performed well in its testing window, which is the out-of-sample period reserved for evaluation.
- Only after that testing period is complete, and after the required freeze period has passed, can the model enter usage.

This is important because many historical backtest can appear stronger than they really are if model selection is influenced, directly or indirectly, by future results. Our process is designed to avoid that problem. By separating training, testing, and usage, and selecting models based only on testing performance, we reduce look-ahead bias, reduce hidden information leakage, and make the resulting historical record more representative of live use. That is a meaningful differentiator in time-series modeling. It reduces the risk of look-ahead bias, hidden information leakage, and overstated historical performance, and it makes the resulting track record more credible as a representation of live use. This is a key reason why we believe our backtests better reflect reality rather than hindsight.

10. Candidate Evaluation and Model Selection

Candidate models are evaluated based on the F1 scores analyzed during the testing windows. Only the top F1 score from each model stage is selected to be a potential candidate for usage. In the specific usage period, there can be up to four transfer learning models eligible to be utilized in live score production.

Out of the four eligible transfer-learning candidates available, the two highest-ranked candidates by F1 score that satisfy the minimum performance threshold are selected for daily scoring. A minimum F1 score of 0.50 applies for deployment. Models below that level are not included in active scoring.

If only one transfer-learning candidate meets the threshold, that model may be used alone. If no transfer-learning candidate meets the threshold, the methodology may fall back to an eligible base model for the relevant usage period. If no eligible model meets the threshold, the feature set is reviewed during the model freeze period. Prior models that were eligible will continue to be utilized until the feature set is updated and a model with an eligible F1 score is available. All feature set changes, and model retrains for this specific scenario, are logged and audited.

[3] Data leakage occurs when information that would not have been available at the decision time (e.g., future observations, labels, or data revised after the fact) is inadvertently included in model training, feature construction, preprocessing, or model selection. Leakage can happen through look-ahead bias, using statistics computed over the full sample (such as normalization that includes the test period), or any workflow that allows test/usage-period information to influence the training process. The result is an overstated backtest and a model that is unlikely to perform similarly in live use.

11. Daily Probability Estimation

Eligible models are run each day on the current feature set in production mode. Each model produces three probabilities representing the assessed likelihood that the covered exposure is undervalued, fair-valued, or overvalued over the decision horizon.

Daily scoring follows standardized production output conventions so that the same score-construction logic can be applied consistently across model vintages.

12. DLR-Score Construction

The official daily output is the Deep Learning Risk Score ("DLR-Score").

Deep Learning Risk Score = Probability of overvaluation - Probability of undervaluation

This measure expresses the balance between the model's assessed probability of overvaluation and undervaluation. The score ranges from -1 to +1. Lower values indicate that the model assigns more weight to undervaluation than overvaluation. Higher values indicate the opposite. Values near zero indicate relative balance or limited differentiation between the two states.

Where two active models are selected under the eligibility and threshold rules above, the daily DLR-Score for the covered exposure is the arithmetic average of the two model-level scores. This averaging is intended to reduce dependence on a single model vintage while preserving a standardized and transparent daily signal.

To support score stability, the final published DLR output is based on a 10-day exponentially weighted moving average of daily DLR-Scores. This reduces sensitivity to model-replacement effects and unusually large single-day score movements. The same smoothing framework is applied consistently across asset classes:

$$\tilde{x}_t = \sum_{i=0}^{9} w_i x_{t-i}, \qquad w_i = \frac{0.8^i}{\sum_{j=0}^{9} 0.8^j}$$

This smooths out potential volatility of our scores from model replacements or large, single day swings of our DLR-score. This weighting mechanism is maintained throughout each asset class universally.

13. Relative Comparison

DLR-Scores are designed to be comparable across eligible assets or thematic sleeves within a defined rotation universe. The methodology uses those scores as standardized decision inputs for relative ranking rather than discretionary narrative judgments. In simple terms, the methodology is designed to make DLR-Scores comparable across eligible assets by calibrating each asset's score against its own historical move profile, as described in Section 6 above. This makes the role of the DLR-Score consistent: it provides a governed, comparable measure of the model's assessed balance between undervaluation and overvaluation for each covered exposure based on the asset's own threshold.

14. Model Refresh and Replacement

Model refresh occurs on a defined quarterly cadence. New base and transfer-learning candidates are trained, evaluated on forward test windows, and activated only when their scheduled usage periods begin and the relevant selection criteria have been met.

As earlier model usage periods expire, replacement occurs according to the predefined schedule and selection rules. This is intended to maintain continuity in live scoring while allowing the methodology to incorporate more recent information as market conditions change.

Model replacement can result in noticeable changes in the underlying DLR-Score because newer model vintages may interpret the available data differently from older vintages. This is an expected consequence of a governed refresh process that is designed to incorporate more recent information and maintain model relevance through time. The methodology addresses this tradeoff through scheduled replacement, score smoothing, and consistent selection rules.

15. Feature Set Evolution and Forward-Only Changes

Feature sets may be updated when new data becomes economically relevant or when a revision improves methodological robustness. For example, newly available ETF market-structure data, such as Bitcoin ETF volume when analyzing Bitcoin, may become informative for certain sleeves after that data begins to exist.

Such changes are applied prospectively only. Historical model records are not rewritten. Previously generated backtest records are not retrospectively altered, and established model lineage remains intact. This forward-only change discipline preserves audit trails and reduces the risk of retroactive optimization of reported histories.

16. Governance, Exception Handling, Error Handling, Auditability, and Reproducibility

The methodology is designed for controlled operations under explicit logging and retained model records.

Material methodology changes are logged, reviewed, and implemented prospectively only.

Control elements include:

- Per-asset model records with stage, period, testing, usage, and F1 metadata.
- Retained model, scaling, architecture, and feature-selection records.
- Deterministic run configuration and structured logging.
- Daily score construction with threshold and window gating.
- Audit outputs for data-quality and date-source diagnostics.

The process is systematic. Once training and deployment cycles begin, model progression is governed by predefined rules and schedule controls rather than day-to-day discretionary intervention.

If required data becomes unavailable or materially delayed for a given day, the process carries forward the most recently available valid data point for that field. This is intended to preserve continuity while maintaining point-in-time discipline.

Errors are documented and reviewed. Where an error results in a change to outputs, the record documents the issue, its materiality, the corrective action taken, and any changes required to prior data points or historical outputs, including a comparison of pre-fix and post-fix results.

17. Publication Timing and Client Delivery

DLR-Scores and underlying index weights are calculated shortly after the NYSE market close and released to clients that evening, before the next market open. We aim to provide outputs within one hour of market close. Per client requests, we can provide intra-day results as well.

If delivery is delayed by more than four hours, clients are notified of the delay.

18. Client-Specific Implementation Parameters

The core DLR-Score methodology is separate from client-specific portfolio implementation. Model outputs are separable from portfolio implementation choices. Client-specific parameters may include:

- Investment universe definition.
- Concentration limits.
- Rebalance cadence, buffer, hysteresis and turnover constraints.
- Additional mandate-specific risk controls.

Accordingly, the same DLR-Score process can support multiple implementation policies and indices, without changing core model governance.

19. Key Methodological Principles and Conclusion

The Thesis Machine Learning framework is built around three principles: point-in-time integrity, walk-forward discipline, and operational reproducibility.

Point-in-time integrity is enforced through date-aware data construction and conservative joining logic. Walk-forward discipline is enforced through strict temporal separation of training, testing, and usage windows. Operational reproducibility is supported through retained model records, governance controls, and rule-based model eligibility.

Within this framework, recent control updates are best understood as strengthening explicit safeguards and audit transparency. They improve process defensibility while remaining consistent with the established methodological intent.

Disclosures

This document is provided for informational purposes only and is intended for institutional or professional recipients. It does not constitute investment advice, legal advice, tax advice, or a recommendation to buy or sell any security or to implement any specific investment strategy.

The methodology described herein is rules-based and designed to operate using information that would reasonably have been available at the relevant time; however, no methodology can eliminate all model risk, data risk, operational risk, or market risk. Outputs may be affected by data limitations, delayed or revised source information, structural market changes, or conditions not fully represented in the historical record.

Any backtested, hypothetical, or model-derived results are presented for illustrative and analytical purposes only. They do not represent actual trading, do not reflect all real-world implementation constraints, and are not guarantees of future performance. Model selection and eligibility are based on predefined testing and governance procedures, but strong historical testing results do not ensure favorable live results.

Methodology changes are logged, reviewed, and implemented prospectively only. Errors, material corrections, and any required output revisions are documented in accordance with the methodology's governance procedures. Third-party data is believed to be reliable but is not guaranteed as to accuracy or completeness.

This document does not imply that any regulator has approved or reviewed the methodology, any related calculations, or any associated performance presentation.