**Report: Solar Forecasting – Progress Overview**

**Κεφάλαιο 1: Βιβλιογραφική Ανασκόπηση**

Η πρόβλεψη ηλιακής ενέργειας (solar forecasting) αποτελεί ένα διαρκώς αναπτυσσόμενο πεδίο έρευνας, κρίσιμο για τη σταθερότητα και αποδοτικότητα των ενεργειακών δικτύων. Το πρόβλημα αυτό μπορεί να προσεγγιστεί από διάφορες οπτικές γωνίες, με έμφαση σε:

* **Statistical Models:** Όπως τα ARIMA και SARIMA, που βασίζονται σε υποθέσεις γραμμικότητας και stationarity. Χρησιμοποιούνται κυρίως για univariate προβλέψεις μικρού ορίζοντα.
* **Machine Learning (ML):** Περιλαμβάνουν τα Support Vector Machines (SVM), Decision Trees, Random Forest και ιδιαίτερα τα Gradient Boosted Trees (XGBoost, LightGBM), που διαχειρίζονται καλά multivariate structured δεδομένα και επιτρέπουν explainability.
* **Deep Learning (DL):** Μοντέλα όπως τα LSTM, GRU, CNN-LSTM και Encoder-Decoder αρχιτεκτονικές, που έχουν αποδειχθεί πολύ ισχυρά σε χρονοσειρές λόγω της ικανότητάς τους να μαθαίνουν μακροχρόνιες εξαρτήσεις.
* **Transformer-based approaches:** Νεότερες μέθοδοι, όπως Temporal Fusion Transformer (TFT), Informer, Autoformer κ.ά., με δυνατότητες multi-horizon prediction και ερμηνευσιμότητας μέσω attention μηχανισμών.
* **Hybrid models:** Συνδυάζουν DL και ML, π.χ. LSTM για feature extraction + XGBoost για τελική πρόβλεψη. Εμφανίζονται όλο και συχνότερα σε papers λόγω των πλεονεκτημάτων και των δύο κόσμων.

Επίσης, αναγνωρίζεται η σημασία των **Numerical Weather Prediction (NWP)** δεδομένων, καθώς και των ημερολογιακών/γεωγραφικών χαρακτηριστικών για την ακρίβεια της πρόβλεψης.

Πρόσφατες εργασίες έχουν εστιάσει σε προσαρμοστικές αρχιτεκτονικές, μεταφορά μάθησης (transfer learning), αλλά και σε explainable AI (π.χ. SHAP) ώστε να ερμηνεύεται καλύτερα η πρόβλεψη παραγωγής.

**Ενδεικτικά State of the Art μοντέλα:**

1. **CNN-LSTM** – Συνδυασμός Convolutional layers (για εντοπισμό χρονικών patterns) και LSTM για long-term dependencies. Πολύ αποδοτικό σε multivariate forecasting.
2. **XGBoost** – Ένα από τα ισχυρότερα μηχανικά μοντέλα πρόβλεψης για structured data. Πολύ υψηλή ακρίβεια και σταθερότητα.
3. **Temporal Fusion Transformer (TFT)** – Νεότερη DL αρχιτεκτονική που χρησιμοποιεί attention και εισάγει interpretability.
4. **N-BEATS** – Deep feedforward αρχιτεκτονική ειδικά σχεδιασμένη για πρόβλεψη χρονοσειρών, χωρίς recurrence.
5. **Seq2Seq με Attention** – Encoder-decoder δομή με attention μηχανισμό, συχνά εφαρμόζεται σε forecasting με πολύπλοκες εξαρτήσεις.

**Κεφάλαιο 2: Ανάλυση Δεδομένων**

Τα δεδομένα προέρχονται από το **SunDance Dataset**, το οποίο περιλαμβάνει παραγωγή ηλιακής ενέργειας και μετεωρολογικά δεδομένα για 100 sites στη Βόρεια Αμερική για το 2015.

Στα πειράματα που έχουμε τρέξει έχουν χρησιμοποιηθεί τα αρχεία “SunDance\_1007.csv” από τους φακέλους energy και weather του dataset.

* **Διάρκεια:** 1 έτος (2015)
* **Granularity:** 1 ώρα
* **Χαρακτηριστικά:** παραγωγή [kW], θερμοκρασία, υγρασία, ταχύτητα ανέμου, ηλιακή ακτινοβολία και πολλά άλλα (σύνολο 37 χαρακτηριστικά καιρού και 4-8 χαρακτηριστικά ενέργειας)

**Προεπεξεργασία Δεδομένων**

Στα δεδομένα που χρησιμοποιούμε έχουν γίνει οι παρακάτω παρεμβάσεις:

* Συγχώνευση δεδομένων καιρού και ενέργειας (merge\_weather\_energy.py)
* Parsing ημερομηνίας και ώρας από στήλες (dateyear, datemon, datemday, datehour, datemin)
* Δημιουργία timestamp και ρύθμιση ως index
* Στρογγυλοποίηση timestamp στην πλησιέστερη ώρα
* Εσωτερική συγχώνευση (inner join) με βάση κοινό timestamp και ευθυγράμμιση χρονικών σειρών
* Καθαρισμός και οπτική ανάλυση των καιρικών δεδομένων (Preprocess\_weather.py)
* Ανάθεση ονομάτων στηλών λόγω έλλειψης (με βάση το documentation)
* Δημιουργία και χρήση timestamp
* Έλεγχος για NaNs
* Στατιστική ανάλυση
* Plot για κατανομή θερμοκρασίας ανά ώρα
* Καθαρισμός δεδομένων ενέργειας (Preprocess\_data.py)
* Οπτικοποίηση χρονοσειράς παραγωγής (π.χ. use [kW])
* Μετονομασία χρονικών πεδίων
* Έλεγχος για NaN, περιγραφικά στατιστικά
* Sorting και indexing με βάση timestamp
* Δημιουργία νέων χαρακτηριστικών (feature\_engineering.py)
* Χρονικά χαρακτηριστικά: hour, dayofweek, month, is\_weekend, is\_daylight
* Lag χαρακτηριστικά: generation\_lag\_1h, 2h, 3h, 24h
* Rolling μέσοι όροι (π.χ. 3h rolling μέσος για generation και tempm)
* Αφαίρεση NaNs που προέκυψαν από τα παραπάνω

**Κεφάλαιο 3: Μοντέλα και Πρόταση**

**Α. State-of-the-Art Μοντέλα που χρησιμοποιήθηκαν:**

1. **CNN-LSTM (DL):** Συνδυασμός 1D Convolution Νευρωνικό (για local patterns) και LSTM (για long-term dependencies). Κατάλληλο για πολυπαραγοντικά χρονικά δεδομένα.
2. **XGBoost (ML):** Gradient Boosted Trees. Πολύ αποδοτικό για προβλήματα structured data και forecasting με εξωτερικά χαρακτηριστικά (lags, καιρικά κ.ά.).
3. **Random Forest (ML):** Ensemble από decision trees, σταθερό και απλό σε ερμηνεία. Δεν υποστηρίζει εγγενώς time dependencies.
4. **NWP + XGBoost (ML**): Πρόβλεψη βασισμένη αποκλειστικά σε Numerical Weather Prediction χαρακτηριστικά με χρήση XGBoost ως regressor.
5. **Hybrid CNN → LSTM → XGBoost (DL+ML):** Χρήση CNN και LSTM ως feature extractors, με XGBoost στο τέλος ως regressor. Deep Feature Fusion αρχιτεκτονική.
6. **Hybrid LSTM → XGBoost (DL+ML):** LSTM ως encoder για deep features και XGBoost για τελική πρόβλεψη.
7. **Hybrid LSTM → Random Forest (DL+ML):** Παρόμοιο με το παραπάνω αλλά με Random Forest στην έξοδο.

**Β. Άλλα μοντέλα που χρησιμοποιήθηκαν:**

1. **ARIMA (Statistical/ML):** Γραμμικό στατιστικό μοντέλο πρόβλεψης, κατάλληλο για univariate χρονοσειρές. Δεν ενσωματώνει εξωτερικά χαρακτηριστικά.
2. **LSTM (DL):** Recurrent νευρωνικό δίκτυο που μαθαίνει διαχρονικές εξαρτήσεις. Αποτελεί κλασική λύση για univariate και multivariate χρονοσειρές.

Όλα τα παραπάνω μοντέλα ανήκουν σε κατηγορίες Deep Learning (DL), Machine Learning (ML) ή Στατιστικής Μοντελοποίησης (ARIMA).

**Γ. Πλαίσια και Βιβλιοθήκες:** Χρησιμοποιήθηκαν εργαλεία όπως TensorFlow, Keras, scikit-learn και XGBoost.

**!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!**

**Η πρόταση:** Παρόλο που αρχιτεκτονικές τύπου **LSTM → XGBoost** έχουν χρησιμοποιηθεί στη βιβλιογραφία, επιλέγονται εδώ λόγω της απόδοσης και ευελιξίας τους. Ωστόσο, η εργασία μπορεί να εμπλουτιστεί στο μέλλον με **προηγμένα attention-based μοντέλα** (π.χ. Transformer-based feature extractors ή multi-task encoders), ώστε να ενισχυθεί η γενίκευση και να βελτιωθεί η αξιοπιστία σε δύσκολες συνθήκες πρόβλεψης.

!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!

**Κεφάλαιο 4: Σύγκριση Αποτελεσμάτων**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Μοντέλο** | **RMSE** | **MAE** | **R² Score** | **Σχολιασμός** |
| **XGBoost** | 0.84 | 0.44 | 0.9068 | Καλύτερη συνολική απόδοση. Χρησιμοποιεί lags, weather & time features |
| **Random Forest** | 0.85 | 0.47 | 0.9046 | Πολύ κοντά σε XGBoost, ελαφρώς πιο ομαλή πρόβλεψη |
| **LSTM** | 1.23 | ~0.65 | ~0.85 | Μαθαίνει χρονικά μοτίβα, χρειάζεται tuning για peaks |
| **CNN-LSTM** | 1.34 | ~0.70 | ~0.83 | Λεία curves, oversmoothing – καλή γενική τάση |
| **CNN → LSTM → XGBoost** | 1.31 | 0.71 | 0.7721 | Deep Feature Fusion, αλλά όχι καλύτερο από XGBoost |
| **Hybrid LSTM → RandomForest** | 1.38 | 0.78 | 0.7490 | Περιορισμένη μάθηση από το LSTM – overshooting |
| **Hybrid LSTM → XGBoost** | 1.25 | 0.70 | 0.7939 | Βελτίωση σε σχέση με RF, αλλά κάτω από XGBoost full |
| **NWP + XGBoost only** | 1.43 | 0.88 | 0.7295 | Χρησιμοποιεί μόνο καιρικά δεδομένα ➝ baseline χωρίς lags |
| **ARIMA** | 2.80 | 2.42 | -0.0355 | Πολύ κακή επίδοση – δεν εκμεταλλεύεται εξωτερικά χαρακτηριστικά |

**Συμπέρασμα:**

Το μοντέλο που σημείωσε την καλύτερη απόδοση με τα συγκεκριμένα δεδομένα όπως φαίνεται και παραπάνω είναι το XGBoost. Κάτι που δικαιολογείται από το γεγονός ότι λειτουργεί καλύτερα και με μεγαλύτερη ακρίβεια στις short term προβλέψεις**.**