



RAPPORT DE PROJET DEEP LEARNING

“ ANALYSE ET PRÉDICTION DE TRAFIC RÉSEAU 5G ”

Étudiant : NACHDA NOUROUDDINE SOIBAHA

Encadrant : Pr. OKAR CHAFIK

Formation : 4ème année Systèmes de Télécommunications & Réseaux

Année universitaire : 2025/2026

Table des Matières

1. INTRODUCTION
2. EXPLORATION ET ANALYSE DES DONNÉES
3. MÉTHODOLOGIE ET EXPÉRIMENTATION
4. RÉSULTATS ET ANALYSE
5. DISCUSSION
6. CONCLUSION ET PERSPECTIVES
7. BIBLIOGRAPHIE ET ANNEXES

1. Introduction

1.1 Contexte

L'avènement de la 5G a révolutionné les communications mobiles en offrant des débits élevés (jusqu'à 10 Gbps), une latence ultra-faible ($<1\text{ms}$) et une capacité accrue. Cependant, la **gestion intelligente du trafic réseau** reste un défi majeur pour les opérateurs réseau pour garantir une **qualité de service (QoS)** optimale, notamment pour les applications critiques comme la visioconférence, le gaming en ligne et le streaming. La capacité à anticiper le comportement du trafic réseau permettrait une allocation proactive des ressources et une optimisation dynamique, garantissant ainsi une expérience utilisateur fluide et stable.

1.2 Problématique

Les applications de visioconférence, comme Microsoft Teams, présentent des caractéristiques de trafic complexes : bidirectionnalité, sensibilité extrême à la latence et au jitter, et patterns de trafic variables selon l'activité (audio, vidéo,

partage d'écran). **Comment pouvons-nous anticiper précisément le comportement futur du réseau pour ces applications critiques ?** Une bonne prédiction du débit et du nombre de paquets permettrait une gestion proactive du réseau, mais cette tâche est rendue difficile par la nature bruyante et volatile du trafic réseau réel.

1.3 Objectifs du Projet

Ce projet vise à répondre à cette problématique à travers les objectifs suivants :

- Prédire simultanément le débit (**throughput**) et le nombre de paquets (**packet_count**) pour les 10 prochaines secondes sur un flux Microsoft Teams
- Mener une étude comparative exhaustive de 23 modèles répartis en 3 catégories : **modèles de référence (baselines)**, **deep learning** et **ensembles**
- Identifier l'architecture optimale pour la prédiction multi-variables de trafic 5G
- Analyser la valeur diagnostique du couple throughput/packet_count pour le dépannage (troubleshooting) réseau

1.4 Jeu de Données

Le projet utilise le **5G Traffic Datasets** disponible sur Kaggle, contenant des traces réelles collectées entre mai et octobre 2022 sur le réseau d'un opérateur mobile sud-coréen majeur. Les données ont été capturées via un terminal Samsung Galaxy A90 5G équipé d'un modem Qualcomm Snapdragon X50, utilisant l'application PCAPdroid.

Caractéristiques principales :

- Volume total : ~45 GB (75 fichiers CSV)
- Durée totale : 328 heures de collecte
- Granularité : Paquet réseau (niveau milliseconde)
- Applications : 6 catégories couvrant les usages 5G majeurs

Pour ce projet, nous nous concentrons sur le fichier MS_Teams_1.csv (1.1 GB, 4 millions de paquets, 2h46 de session) de la catégorie "Video Conferencing".

2. Exploration et Analyse des Données

2.1 Prétraitement et Feature Engineering

Les données brutes au niveau paquet (4+ millions d'observations) sont trop granulaires pour une modélisation directe. Nous avons donc effectué un resampling temporel pour agréger les données par seconde :

Variables cibles calculées :

- Throughput (Mbps) = $(\sum \text{Length des paquets sur 1s} \times 8 \text{ bits/byte}) / 1,000,000$
- Packet Count = Nombre de paquets reçus pendant 1 seconde

Features sélectionnées :

1. `throughput_mbps` - Débit réseau (target + feature)
2. `packet_count` - Nombre de paquets (target + feature)
3. `avg_packet_size` - Taille moyenne des paquets
4. `std_packet_size` - Écart-type de la taille des paquets (indicateur de variabilité)

2.2 Analyse Exploratoire (EDA)

Stabilité et Stationnarité

Les séries temporelles montrent une stabilité remarquable :

- Throughput : 2.00 ± 0.17 Mbps (CV = 8.5%)

- Packet Count : 400 ± 28 p/s (CV = 7%)

Le test de stationnarité Augmented Dickey-Fuller confirme que les deux variables cibles sont stationnaires (p-value = 0.0000), éliminant le besoin de différenciation.

Relation entre Variables Cibles

Une corrélation linéaire forte ($r = 0.791$) existe entre le throughput et le packet count, justifiant scientifiquement l'approche multi-output. Cette synergie permet aux modèles d'exploiter les relations entre les métriques.

Analyse de la Saisonnalité et Autocorrélation

La décomposition saisonnière révèle une saisonnalité négligeable (force = 0.006), typique des applications de visioconférence qui n'ont pas de cycles jour/nuit comme le trafic web traditionnel. L'analyse ACF/PACF montre une mémoire temporelle significative sur ~10-15 secondes, guidant le choix de la longueur de fenêtre.

Détection des Patterns et Scénarios

Le clustering K-Means (k=3) identifie trois comportements distincts :

- Cluster 1 (59.1%) : Activité normale basse (1.95 Mbps, 386 p/s)

- Cluster 2 (40.4%) : Activité normale haute (2.09 Mbps, 422 p/s)
- Cluster 3 (0.5%) : Anomalies réseau (0.70 Mbps, 179 p/s)

L'analyse des scénarios réseau révèle que 86.2% du temps correspond à une activité normale, tandis que 0.6% montre des patterns anormaux (faible débit, pics d'activité) ayant une valeur diagnostique importante.

Analyse CRITIQUE du Bruit

Les données présentent un niveau de bruit élevé :

- Throughput : SNR = 0.61 (FAIBLE)
- Packet Count : SNR = 0.92 (FAIBLE)
- Variabilité court terme > long terme (ratio > 1.4)

Cette caractéristique a des implications majeures pour la modélisation, nécessitant une régularisation renforcée, un dropout accru et des fenêtres temporelles optimisées.

3. Méthodologie et Expérimentations

3.1 Formulation du Problème

- Tâche : Prédiction multi-output et multi-step
- Input : Séquence de 60 secondes \times 4 features
- Output : Prédiction des 10 secondes futures \times 2 targets
- Approche : Apprentissage supervisé avec validation temporelle

3.2 Préparation des Données

- Split temporel : 70% train / 15% validation / 15% test
- Normalisation : StandardScaler (moyenne=0, écart-type=1)
- Séquences : Format 3D pour le Deep Learning (batch, timesteps, features)
- Justification fenêtre 60s : Optimisée empiriquement pour capturer 3 changements de régime (1 changement/20s) sans surapprendre le bruit

3.3 Architecture des Modèles Testés

3.3.1 Modèles de Référence (8 modèles)

- Persistence : $\hat{y}(t+1:t+10) = y(t)$ (baseline naïve)
- Moving Average : Moyenne glissante optimisée (fenêtre=5s)
- Exponential Smoothing : Lissage exponentiel avec trend
- ARIMA : AutoRegressive Integrated Moving Average
- VAR : Vector AutoRegression (exploite la corrélation entre targets)
- Dynamic Regression : Régression linéaire avec lag features

- Random Forest : Ensemble d'arbres avec MultiOutputRegressor
- XGBoost : Gradient boosting optimisé

3.3.2 Deep Learning (6 modèles)

- MLP : Perceptron multi-couches (baseline DL)
- LSTM : Long Short-Term Memory (récurrent standard)
- GRU : Gated Recurrent Unit (alternative légère à LSTM)
- CNN-LSTM : Architecture hybride (patterns locaux + dépendances temporelles)
- Transformer : Mécanisme d'attention
- BiLSTM : LSTM bidirectionnel

Optimisations pour données bruitées :

- Dropout augmenté (0.3-0.4)
- Régularisation L2
- Early Stopping agressif (patience=8)
- Learning Rate Scheduling adaptatif

3.3.3 Ensemble Learning (9 modèles)

- Simple Averaging : Moyenne arithmétique des prédictions
- Weighted Average : Pondération basée sur les performances (1/MAE)

- Stacked Generalization : Meta-learner (Ridge) apprenant la combinaison optimale

3.4 Métriques d'Évaluation

- MAE (Mean Absolute Error) : Erreur moyenne absolue (moins sensible aux outliers)
- R^2 (Coefficient de Détermination) : Proportion de variance expliquée
- Focus : Métriques globales (moyenne des deux targets) pour comparaison équitable

4. Résultats et Analyse

4.1 Performance Comparative

Classement Final (Top 10 sur 23 modèles) :

Rang	Modèle	Catégorie	MAE	R^2
1	Stacked Generalization	Ensemble	0.1761	0.5697

2	LSTM	Deep Learning	0.2158	0.4206
3	BiLSTM	Deep Learning	0.2234	0.3388
4	GRU	Deep Learning	0.2244	0.3897
5	Weighted Average	Ensemble	0.2352	0.3887
6	CNN-LSTM	Deep Learning	0.2384	0.3406
7	Simple Average	Ensemble	0.2398	0.3809
8	XGBoost	Baseline	0.2427	0.3636
9	Random Forest	Baseline	0.2569	0.3216
10	Dynamic Regression	Baseline	0.2327	0.2404

4.2 Analyse par Catégorie

4.2.1 Modèles de Référence

Les modèles traditionnels montrent des performances variables :

- XGBoost émerge comme le meilleur baseline (MAE=0.2200), exploitant efficacement les interactions complexes entre les 240 features
- Random Forest et Dynamic Regression offrent des performances compétitives
- Les méthodes statistiques simples (Persistence, ARIMA) échouent complètement (R^2 négatifs), incapables de capturer la dynamique complexe du trafic

4.2.2 Deep Learning

La LSTM standard surpasse toutes les autres architectures neuronales :

- Avantages : Mémoire cellulaire, gates filtrant le bruit, équilibre paramètres/performance
- GRU performe légèrement moins bien (2 gates vs 3), mais plus rapide à l'entraînement
- CNN-LSTM capte bien les patterns locaux mais perd en performance globale
- Transformer échoue dramatiquement (sur-paramétrage, manque de données, amplification du bruit)
- BiLSTM déçoit (inutile en prédiction temps réel où le futur n'est pas disponible)

4.2.3 Ensemble Learning - RÉSULTAT MAJEUR

Le Stacked Generalization achieve des performances exceptionnelles :

- Amélioration de +18.4% vs le meilleur modèle individuel (LSTM)
- Meta-learner Ridge apprend dynamiquement quand faire confiance à chaque modèle
- Analyse des coefficients :
 - CNN-LSTM (0.2715) : Expert des patterns courts
 - Transformer (0.2797) : Gère les cas complexes
 - BiLSTM (0.2404) : Expert temporel long

4.3 Visualisations Clés

Courbes de Prédiction

Les prédictions du modèle Stacking suivent fidèlement les valeurs réelles, avec une réduction notable du bruit par rapport aux modèles individuels. La courbe montre une capacité à anticiper les tendances générales tout en lissant les fluctuations excessives.

Distribution des Erreurs

La distribution des erreurs du modèle champion est plus étroite et centrée autour de zéro, indiquant à la fois une meilleure précision et l'absence de biais systématique.

5. Discussion

5.1 Synthèse des Résultats

Ce projet démontre de manière convaincante que l'approche ensembliste par Stacked Generalization est optimale pour la prédiction multi-variables de trafic 5G. En combinant judicieusement les forces de modèles diversifiés (MLP, LSTM, CNN-LSTM, Transformer), le meta-learner achève une performance inégalée, surpassant même la meilleure architecture individuelle de 18.4%.

5.2 Contributions Clés

1. Validation de l'approche multi-output : La forte corrélation (0.791) entre throughput et packet count est effectivement exploitable pour améliorer les prédictions
2. Optimisation de l'architecture : La LSTM simple, avec régularisation appropriée, est l'architecture neuronale la plus efficace pour ces données bruitées

3. Démonstration de la puissance du stacking : L'apprentissage de la combinaison optimale de modèles est supérieur à toute architecture monolithique
4. Analyse du bruit : La caractérisation approfondie du bruit a guidé des choix d'optimisation critiques

5.3 Limites et Défis

- Données très bruitées : Le faible SNR (0.61-0.92) limite les performances absolues atteignables
- Volume de données : 10,007 points sont suffisants mais limitent les architectures complexes comme le Transformer
- Généralisation : Les résultats sont spécifiques à Microsoft Teams et nécessiteraient une validation sur d'autres applications
- Temps de calcul : Certains modèles (BiLSTM, Stacking) ont des temps d'inférence élevés

5.4 Perspectives et Travaux Futurs

- Extension multi-applications : Tester la généralisation sur Zoom, Google Meet, Netflix
- Features externes : Incorporer des métriques radio (RSRP, RSRQ, SINR)

- Débruitage avancé : Explorer des techniques de filtrage adaptatif
- Architectures légères : Développer des modèles optimisés pour le déploiement en temps réel
- Online Learning : Implémenter des mécanismes d'adaptation en continu aux changements de patterns réseau

6. Conclusion

Ce projet a abordé le challenge complexe de la prédiction multi-variables de trafic 5G pour Microsoft Teams à travers une méthodologie rigoureuse et une expérimentation exhaustive.

Principales réalisations :

- Prétraitement et analyse approfondie d'un dataset réel de trafic 5G
- Implémentation et évaluation comparative de 23 modèles à travers 3 catégories
- Démonstration de la supériorité de l'approche ensembliste par Stacked Generalization

- Obtention d'une amélioration de performance de 18.4% vs le meilleur modèle individuel

Conclusion principale : Le Stacked Generalization émerge comme la stratégie optimale, avec un MAE de 0.1761 et un R^2 de 0.5697 - des performances remarquables compte tenu du caractère très bruité des données réelles.

Ces résultats ouvrent la voie à des systèmes de gestion proactive de réseau 5G capables d'anticiper les besoins en ressources et de garantir une qualité de service optimale pour les applications critiques comme la visioconférence.

7. Bibliographie et Annexes

7.1 Références

- Kaggle : 5G Traffic Datasets :
[https://www.kaggle.com/datasets/kimdaegyeom/5g-traffic-datasets?select=5G Traffic Datasets](https://www.kaggle.com/datasets/kimdaegyeom/5g-traffic-datasets?select=5G+Traffic+Datasets)
- TensorFlow/Keras Documentation
- Scikit-learn Documentation
- Brownlee, J. "Deep Learning for Time Series Forecasting"

7.2 Annexes

- Code source complet (notebook Jupyter)
- Visualisations supplémentaires
- Métriques détaillées par modèle
- Architecture détaillée des modèles Deep Learning