2025-04-26 14:28

Tags: [[DeepLearning]], [[Machine Learning]]

# Early wldfire detection using different machine learning algorithms

* **Abstract summary**: Nghiên cứu này đề xuất một hệ thống phát hiện cháy rừng sớm bằng cách sử dụng các nút cảm biến không dây giá rẻ kết hợp với các phương pháp AI detection. Hệ thống gồm các cảm biến đo nhiệt độ, độ ẩm, khói và một module truyền thông không dây. Bốn thuật toán máy học được đánh giá *(bao gồm decision trees, random forests, SVM và KNN),* trong đó **Random Forest cho kết quả chính xác cao nhất với độ chính xác 77.95%.** Hệ thống này hiệu quả và tiết kiệm chi phí, phù hợp cho việc giám sát cháy rừng trên diện rộng.

### 1. Introduction

* Trong những năm gần đây, cháy rừng ngày càng xảy ra nhiều và nghiêm trọng hơn do biến đổi khí hậu, thay đổi mục đích sử dụng đất và các hoạt động của con người.
* Biến đổi khí hậu làm tăng nhiệt độ, thay đổi lượng mưa, khiến đất đai khô hạn hơn, dễ cháy hơn. Hoạt động của con người như đốt phá rừng, đốt lửa trại, bắn pháo hoa cũng làm tăng nguy cơ cháy.
* Cháy rừng gây ra nhiều hậu quả:
  + phá hủy môi trường sống, mất đa dạng sinh học, xói mòn đất, ảnh hưởng xấu đến sức khỏe con người và động vật qua khói bụi, gây thiệt hại kinh tế lớn.
  + Chi phí dập tắt cháy rừng ở Mỹ năm 2020 vượt hơn 2,7 tỷ USD. Vì vậy, cần có các chiến lược phát hiện và giảm thiểu cháy rừng hiệu quả.

### 2.Design of early wildfire detection systems

* **Các cách phát hiện cháy rừng phổ biến:**
  + **Vệ tinh (satellite-based):** Bao phủ diện rộng nhưng cập nhật dữ liệu chậm.
  + **UAVs (máy bay không người lái):** Cho hình ảnh chất lượng cao, nhưng hạn chế về thời gian bay và vùng phủ sóng.
  + **Mạng cảm biến mặt đất (ground-based sensor netwoks):** Theo dõi liên tục, triển khai được ở những điểm then chốt.
* **Vấn đề với các cách bố trí cũ:**
  + **Dạng đường thẳng hoặc lưới:** Hoạt động kém hiệu quả trong địa hình phức tạp như núi đồi, rừng rậm.
* **Ý nghĩa trong nghiên cứu:** Nghiên cứu này sử dụng mạng cảm biến mặt đất để tận dụng ưu điểm giám sát liên tục, cải thiện khả năng phát hiện cháy sớm trong môi trường thực tế.

### 3. Satellite-based monitoring

* Có độ phủ sóng rất lớn, phù hợp để phát hiện các đám cháy rừng diện rộng
* Tuy nhiên *temporal-resolution hạn chế* (nghĩa là thời gian giữa các lần thu thập dữ liệu khá dài, không liên tục, không cập nhật thường xuyên) khiến cho việc phát hiện bị chậm trệ
* Các nghiên cứu gần đây đã giúp cải thiện temporal-resolutin và độ chính xác của hệ thống này, khiến việc early detection đã dễ dàng hơn

### 4. UAV-based systems

* Có khả năng capture được hình và video *chất lượng cao*, sau đó có thể áp dụng các thuật toán học máy để giúp phát hiện cháy rừng.
* Có tính linh hoạt cao và có thể deploy dễ dàng. nhưng tầm hoạt đông và thời gian hoạt động hạn chế

### 5. Ground-based sensor networks

* **Mạng cảm biến không dây (WSN)** đang ngày càng được sử dụng nhiều để phát hiện cháy rừng vì khả năng giám sát liên tục theo thời gian thực. Trong nghiên cứu này, cảm biến được bố trí theo **hình tam giác** nhằm tối ưu hóa độ chính xác và độ phân giải không gian, đồng thời tăng khả năng chịu lỗi khi một nút cảm biến bị hỏng.
* **Phát hiện cháy sớm** rất quan trọng để kiểm soát cháy hiệu quả. Các phương pháp truyền thống (thủ công) thường chậm và thiếu chính xác. Vì thế, **machine learning** và **dữ liệu viễn thám** (như ảnh vệ tinh, UAV) đang được áp dụng rộng rãi để cải thiện khả năng phát hiện.
* Các nghiên cứu trước đây đã dùng nhiều thuật toán như **Random Forest, SVM, CNN, Artificial Neural Networks** để nhận diện cháy, cho kết quả khả quan.
* **WSN kết hợp UAV** và các thuật toán thông minh như **fuzzy logic** cũng được phát triển để theo dõi và cảnh báo cháy kịp thời.
* **Chỉ số thời tiết cháy (Fire Weather Index - FWI)** được dùng để đánh giá nguy cơ cháy dựa trên độ ẩm, gió, nhiệt độ, lượng mưa,… Có hai cách phổ biến: **phương pháp của Canada** và **phương pháp của Hàn Quốc**. Phương pháp của Canada chính xác hơn, nhanh hơn và tiết kiệm năng lượng nên được ưu tiên sử dụng trong nghiên cứu này.
  + **Phương pháp Canada:** Được xây dựng trên 6 yếu tố chính liên quan đến điều kiện dễ xảy ra cháy rừng. Các yếu tố này phản ánh khả năng cháy và mức độ nguy hiểm của lửa trong điều kiện thực tế.
  + **Phương pháp Hàn Quốc:** Lấy nền tảng từ phương pháp Canada nhưng được điều chỉnh để phù hợp với đặc điểm địa lý, khí hậu và thảm thực vật của Hàn Quốc, thêm các yếu tố như địa hình và lớp phủ đất.
* Nghiên cứu nhấn mạnh việc **tối ưu hóa cách lắp đặt cảm biến và thu thập dữ liệu tại chỗ**, nhằm nâng cao hiệu quả của các mô hình machine learning khi phát hiện cháy rừng.

### 6. Architecture of the proposed early detection system

![[Pasted image 20250427104403.png]]

* Ba **nút cảm biến** được lắp đặt, mỗi nút có các cảm biến đo: **nhiệt độ, độ ẩm tương đối, CO và CO₂**.
* Các nút cảm biến được bố trí theo **hình tam giác** để tối ưu hóa độ chính xác thu thập dữ liệu và đảm bảo hệ thống vẫn hoạt động nếu một nút bị hỏng.
* *Training data* cho mô hình machine learning được thu thập bằng cách thay đổi vị trí đám cháy và khoảng cách giữa các nút.
* Các nút được đặt ở **độ cao 1 mét** so với mặt đất, tại các đỉnh trên của một **khối lập phương** có chiều cao 1m.
* Dữ liệu được thu thập:
  + Bắt đầu **30 phút trước khi đốt lửa**.
  + Cứ mỗi **10 giây** ghi một mẫu dữ liệu.
  + Thí nghiệm cháy kéo dài **30 phút**, sau đó tiếp tục thu thêm dữ liệu **30 phút sau khi lửa tắt**.
* Đám cháy thử nghiệm được tạo ra bằng cách đốt **2kg gỗ teak** trong không gian nhỏ (30cm x 30cm).
* Các thông tin khác như **lượng mưa, tốc độ gió** cũng được thu thập từ hệ thống khí tượng.
* Mỗi vụ cháy sẽ tạo ra **3 bộ dữ liệu** từ 3 cảm biến, tất cả dùng để **huấn luyện mô hình machine learning**. **Note**:
  + **Vì sao dùng nhiều cảm biến (nhiệt độ, độ ẩm, CO, CO₂)?**  
    → Đám cháy làm tăng nhiệt độ, thay đổi độ ẩm, sinh ra khí CO và CO₂ — nhờ vậy hệ thống dễ dàng nhận diện đám cháy.
  + **Tại sao bố trí hình tam giác?**  
    → Giúp **khoanh vùng vị trí đám cháy** chính xác hơn (giống như định vị GPS dựa trên nhiều vệ tinh) và tăng độ an toàn nếu 1 nút gặp sự cố.

### 7. FWI calculation

* **FWI (Fire Weather Index)** là một chỉ số quan trọng dùng để **đánh giá nguy cơ cháy rừng** trong một khu vực.
* **FWI được tính toán** dựa trên các yếu tố thời tiết: **nhiệt độ, độ ẩm tương đối, tốc độ gió và lượng mưa**.
* Quá trình tính FWI gồm nhiều bước trung gian:
  + **FFMC**: độ ẩm của lớp vật liệu dễ cháy trên bề mặt (nhạy với độ ẩm và mưa).
  + **DMC**: độ ẩm của lớp đất hữu cơ ngay dưới mặt đất.
  + **DC**: độ ẩm của lớp đất hữu cơ sâu hơn (dài hạn hơn).
  + **ISI**: tốc độ lan truyền ban đầu của đám cháy (dựa trên FFMC + gió).
  + **BUI**: tổng lượng nhiên liệu sẵn có để cháy (dựa trên DMC + DC).
  + Cuối cùng, **FWI** được tính từ ISI và BUI.
* **Ngưỡng FWI = 20** được chọn:
  + FWI > 20 → phân loại thành **có nguy cơ cháy**.
  + FWI ≤ 20 → **không có nguy cơ cháy**.
* **Python** được dùng để viết code tính toán FWI từ dữ liệu của các sensor.
* Sau đó, dữ liệu (đã dán nhãn cháy/không cháy) được dùng để **huấn luyện mô hình machine learning**.
* Hiệu suất các thuật toán ML được đánh giá qua **accuracy, recall, precision và F1 score**.

### 8. Machine learning algorithms

* Bài nghiên cứu áp dụng **4 thuật toán machine learning** để so sánh khả năng dự đoán cháy rừng:
  + *Decision Trees*
  + *Random Forests*
  + *Support Vector Machines (SVMs)*
  + *K-Nearest Neighbors (KNN)*
* **Giới thiệu nhanh về từng thuật toán:**
  + **Decision Tree:** đơn giản, hiệu quả, chia nhỏ dữ liệu bằng cách chọn đặc trưng (feature) tốt nhất để tách.
    - Các **hyperparameters**: độ sâu tối đa của cây, số lượng mẫu tối thiểu để chia nhánh, số lượng mẫu tối thiểu ở lá.
  + **Random Forest:** tập hợp nhiều cây quyết định và tổng hợp kết quả để tăng độ chính xác.
    - Các **hyperparameters**: số lượng cây trong rừng, số lượng đặc trưng được xét khi tìm cách chia tốt nhất.
  + **SVM:** tìm một **hyperplane** phân chia dữ liệu tốt nhất bằng cách **maximizing marigns** giữa các lớp.
    - Các **hyperparameters**: loại hàm kernel (ví dụ linear, RBF), tham số điều chỉnh (regularization C), độ rộng của kernel.
  + **KNN:** phân loại dựa trên **k điểm gần nhất** trong tập huấn luyện.
    - Các **hyperparameters**: số lượng k láng giềng, loại khoảng cách (ví dụ Euclidean).
* Để **chọn mô hình tốt nhất**, nhóm nghiên cứu đã dùng kỹ thuật **cross-validation** và **grid search** để tìm bộ hyperparameters tối ưu và đánh giá hiệu suất mô hình.

### 9. Result and discussion

* **Dữ liệu nghiên cứu** gồm 150.000 mẫu với các thuộc tính như: vĩ độ, kinh độ, năm, tháng, ngày, độ ẩm tương đối, nhiệt độ, lượng mưa hàng ngày, gió và chỉ số FWI (Fire Weather Index).
* **Mối liên hệ nhiệt độ và FWI**:
  + Fig 2 cho thấy có **mối tương quan rõ** giữa nhiệt độ và FWI với hiện tượng cháy (fire) và không cháy (non-fire).
  + **Nhiệt độ và FWI càng cao** → **Khả năng cháy càng lớn**.  
    ![[Pasted image 20250427122017.png]]
* **Phân tích tương quan**:
  + Fig 1A là **correlation heatmap** thể hiện độ tương quan của mối liên hệ giữa các biến số trong tập dữ liệu.  
    ![[Pasted image 20250427122301.png]]
  + **KDE plot** (Fig 3) giúp hình dung **phân bố nhiệt độ** giữa lớp fire và non-fire → lớp fire nghiêng về phía **nhiệt độ cao hơn**.  
    ![[Pasted image 20250427122341.png]]
* **Kết quả các mô hình máy học**:
  + **Decision Tree**:
    - Độ chính xác phân loại cháy chỉ **52.6%**, không cháy **79%**.
  + **SVM**:
    - Hiệu suất rất kém, không phân biệt tốt giữa fire và non-fire.
  + **Random Forest**:
    - Chính xác cao nhất:
      * Fire đúng **65.29%**, non-fire đúng **75.83%**.
    - Khi tối ưu hóa siêu tham số (HPO) bằng **Grid Search**, hiệu suất tăng lên:
      * Fire đúng **77.95%**, AUC đạt **0.82**.
* **Tối ưu hóa Random Forest**:
  + **Grid Search** để chọn các tham số tối ưu như: số lượng cây, độ sâu tối đa, tiêu chí chia nhánh...
  + **Phân tích tầm quan trọng đặc trưng** cho thấy **lượng mưa hàng ngày (daily rain)** là yếu tố ảnh hưởng lớn nhất. ![[Pasted image 20250427123325.png]]
  + sau khi tinh chỉnh tham số, cho kết quả tốt hơn rất nhiều so với các mô hình còn lại (SVM, KNN, Decision Tree). ![[Pasted image 20250427123154.png]]
* **So sánh tổng thể**:
  + Dựa trên **ROC** và **độ chính xác**, Random Forest vượt trội hơn Decision Tree, SVM và KNN.
  + Random Forest xử lý tốt dữ liệu lớn, dữ liệu nhiều chiều, mối quan hệ phi tuyến tính, đồng thời chống được hiện tượng **overfitting** nhờ kỹ thuật tổ hợp nhiều cây quyết định. ![[Pasted image 20250427123220.png]]

### 10. Limitations of the study

Nghiên cứu này chứng minh tiềm năng của việc sử dụng các *thuật toán học máy và* *optimizing sensor configurations* cho việc phát hiện cháy rừng sớm. Tuy nhiên, có một số hạn chế cần lưu ý:

1. **Thiết kế sensors và khoảng cách giữa các sensors:** Những yếu tố này dựa trên các thí nghiệm ban đầu và có thể không phải là tối ưu cho tất cả các loại địa hình và điều kiện môi trường.
2. **Các biến môi trường đo được:** Chỉ đo được nhiệt độ, độ ẩm, khí carbon monoxide (CO) và carbon dioxide (CO2). Các yếu tố khác như tốc độ gió, độ ẩm của đất và loại thực vật có thể cung cấp cái nhìn đầy đủ hơn về điều kiện cháy rừng.

### 11. Conclusion

* **Thách thức chính:** Làm sao để sử dụng ít sensors nhất nhưng vẫn đảm bảo chi phí hợp lý và có thể triển khai một cách hiệu quả trong khu vực rừng rộng lớn để phát hiện cháy sớm.
* **Giải pháp nghiên cứu:** Hệ thống cảm biến không dây kết hợp nhiều cảm biến giá rẻ với phương pháp phát hiện dựa trên AI để phát hiện cháy rừng sớm. Nghiên cứu chứng minh rằng hệ thống này có chi phí hợp lý, giảm số lượng cảm biến cần thiết nhưng vẫn đảm bảo độ chính xác cao.
* **Hiệu quả:** Hệ thống này có thể triển khai trên diện tích rừng rộng lớn một cách kinh tế, nhờ giảm chi phí phần cứng và giảm nhu cầu giám sát thủ công.
* **Máy học:** Các bộ phân loại máy học như *decision trees, random forest, SVMs và k-KNN* được áp dụng để so sánh hiệu quả trong việc phát hiện cháy rừng. Kết quả cho thấy *random forest* có hiệu quả tốt nhất, phân loại đúng 77,95% mẫu cháy. Các chỉ số thống kê (p-value và khoảng tin cậy) cho thấy kết quả có ý nghĩa thống kê và độ tin cậy cao.

# References